Texas Law Review

May 2013 Vol. 91 No. 6 Pages 1283 to 1592

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).

Editorial Offices: *Texas Law Review*
727 East Dean Keeton Street, Austin, Texas 78705
(512) 232-1280 Fax (512) 471-3282
tlr@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. BAADE, 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. LITTWIN, 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. SOKOLOW, 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. KEPPLE, 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-LLOYD 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. MCCLAIN, 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 6 May 2013

* * *

Texas Law Review

Volume 91, Number 6, May 2013

NOTES

Articles

Coercion, Compulsion, and the Medicaid Expansion: A Study in the Doctrine of Unconstitutional Conditions

Mitchell N. Berman*

Introduction

The Supreme Court's feverishly anticipated decision in *National Federation of Independent Business v. Sebelius*[1] ("the Health Care Decision"

* Richard Dale Endowed Chair in Law, Professor of Philosophy, The University of Texas at Austin. For extremely helpful comments, challenges, and suggestions, I thank workshop audiences at the law schools of DePaul University, the University of Michigan, Florida State University, the University of Chicago, Duke University, and the University of Texas. For especially valuable

or "*NFIB*") regarding the constitutionality of the Patient Protection and Affordable Care Act (colloquially known as "Obamacare") produced three main holdings concerning two critical provisions of the Act.[2] The first two holdings concerned the "individual mandate" that requires most Americans to maintain "minimum essential" health insurance. First, a 5–4 majority held that this provision exceeded Congress's power under the Commerce Clause.[3] Second, a different 5–4 majority held that this same mandate, which requires those who fail to secure the minimum required health insurance to pay a tax penalty to the IRS, is a constitutional exercise of Congress's taxing authority.[4] The third holding concerned "the Medicaid expansion," which expanded the class of persons to whom the states must provide Medicaid coverage as a condition for receiving federal funds under the Medicaid program.[5] By the more lopsided margin of 7–2, the Court struck down this provision as an impermissible condition on the provision of federal funds to the states.[6]

Of these three holdings, the third—concerning what is often called Congress's "conditional spending power"—is apt to have the most far-reaching consequences beyond health care. The Court's Commerce Clause ruling was predicated on the fact that, in a majority's estimation, Congress was here imposing an unprecedented affirmative obligation upon individuals to enter commerce rather than, as is customary, regulating behavior that was already commercial.[7] Because Congress could not have been expected to impose many—or any—such affirmative obligations even had the dissenters prevailed on the Commerce Clause issue, this ruling will likely have little future impact. And Congress rarely needs to resort to its taxing power to achieve regulatory ends when it can regulate "directly" on the strength of its

---

contributions—at these events, in conversation, or by means of written comments on prior drafts—I wish to particularly acknowledge David Adelman, Matt Adler, Sam Bagenstos, Joseph Blocher, Oren Bracha, Curt Bradley, Curtis Bridgeman, Sam Buell, I. Glenn Cohen, Lee Fennell, Joey Fishkin, Andrew Gold, John Golden, Daniel Halberstam, Bernard Harcourt, Don Herzog, Scott Hershovitz, Andy Koppelman, Guha Krishnamurthi, Marty Lederman, Sandy Levinson, Dan Markel, Richard McAdams, Richard Primus, Jed Purdy, Garrick Pursley, David Rabban, Larry Sager, Mark Schankerman, Margo Schlanger, Neil Siegel, Stephen Siegel, Charlie Silver, James Spindler, David Strauss, Kevin Toh, and Hannah Wiseman, with apologies to those whom I have overlooked. I am also grateful to Paul Still for timely research assistance.

1. 132 S. Ct. 2566 (2012).

2. Whether there were three main holdings, more, or fewer, could be quibbled with. Those who see fewer would contend that the first main holding I identify—that the "individual mandate" was not a permissible exercise of Congress's commerce power—is better characterized as dicta in light of the Court's determination that that provision was a permissible exercise of Congress's taxing power. Those who see more would elevate to "main holding" status other rulings in the case, such as those concerning the anti-injunction act and severability. For my purposes, nothing turns on these possible disagreements.

3. *Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. at 2593.

4. *Id.* at 2575, 2600 (Roberts, C.J.).

5. *Id.* at 2629 (Ginsburg, J., dissenting in part).

6. *Id.* at 2608 (Roberts, C.J.).

7. *Id.* at 2589–90.

commerce power.[8] So the Court's relatively expansive interpretation of Congress's taxing power is not of great moment going forward precisely because its relatively restrictive interpretation of Congress's commerce power is not. But Congress makes habitual (a critic might even say "profligate") use of its conditional spending power.[9] Accordingly, if, as appears to many, the Court has tightened the restrictions on this power, the implications could be profound.

Unfortunately, of the three holdings, the last is not only the most potentially significant, but also the one supported by the least clear rationale. At first blush, to be sure, the majority's reasoning seems straightforward. The key precedent on which the majority drew, *South Dakota v. Dole*,[10] had announced a four-part test governing Congress's use of its spending power to induce state behavior that Congress could not mandate: the spending program must promote "the general welfare," the condition must be unambiguous, the condition must be related to the national interests that the spending would advance, and the condition may not require state recipients to violate the Constitution themselves.[11] No Justices in *NFIB* expressed concern that the Medicaid expansion violated any of these limitations.

In addition to these four restrictions, however, the *Dole* Court read the Spending Clause to impose limits on Congress's ability to "coerce" the states in ways that it could not directly mandate under its other Article I powers.[12] "[I]n some circumstances," the Court observed, "the financial inducement offered by Congress might be so coercive as to pass the point at which 'pressure turns into compulsion.'"[13] It is this prohibition on coercion or compulsion that, a majority of the Court concluded, doomed the Medicaid expansion.[14] While candidly acknowledging that they could provide no guidance regarding how the line between inducement and compulsion would be assessed going forward, seven Justices nonetheless deemed the conditional offer that the Medicaid expansion embodied impermissibly coercive because it gave states "no choice" but to accept.[15]

That, to repeat, is how things appear at first blush. As is often the case, things look rather less clear on second look. For several reasons, it is uncertain that this "no choice" thesis fully captures the majority's reasoning.

---

8. *Cf. id.* at 2578–79 (stating that Congress uses its taxation power when it cannot directly regulate, and contrasting that with its Commerce Clause powers).

9. *See, e.g.*, Bob Drummond, *Limits on Spending Power Seen as Health Ruling's Legacy*, BLOOMBERG (July 1, 2012), http://www.bloomberg.com/news/2012-07-01/limits-on-spending-power-seen-as-health-ruling-s-legacy.html (stating that Congress has used its conditional spending power in many areas).

10. 483 U.S. 203 (1987).

11. *Id.* at 207–08.

12. *Id.* at 211.

13. *Id.* (citing Steward Mach. Co. v. Davis, 301 U.S. 548, 590 (1937)).

14. Nat'l Fed'n of Indep. Bus. v. Sebelius, 132 S. Ct. 2566, 2604 (2012) (Roberts, C.J.).

15. *Id.* at 2603–04, 2606–07.

Among the most important are these. First, neither opinion that combined to constitute the majority on this question—Chief Justice Roberts's for himself and Justices Breyer and Kagan, and the joint opinion of Justices Scalia, Kennedy, Thomas, and Alito[16]—disputed Justice Ginsburg's observation, dissenting on this point, that it would be constitutionally permissible for Congress to repeal the Medicaid Act in its entirety and then enact a new law that mirrored the preexisting law with the Medicaid expansion.[17] Yet if the states had no choice but to accede to the Medicaid expansion, it is hard to see why they would have any more choice but to accede to this new hypothetical Medicaid Act. Second, several passages from the Roberts opinion hint that the constitutional vice was not exactly that states had no real choice other than to accept, but rather that Congress had an impermissible purpose in crafting this particular conditional proposal.[18]

Given the vast potential significance of the Court's holding on conditional spending and the manifest lack of clarity regarding its rationale, a comprehensive and critical assessment of this holding is urgent. That is the ambition of this Article.

The Article advances many claims, some with conviction, others more tentatively. Ruthlessly simplified, the core theses are these. First, insofar as the majority rested its holding of unconstitutionality on the ground that the amount of funds that a state would lose by not agreeing to the condition was so great as to compel the states to accept, that is a highly dubious rationale. Second, it does not necessarily follow that the Court's bottom-line conclusion was wrong. A more promising rationale for that conclusion would be the one merely hinted at by the Chief Justice: Congress's threat to withhold all Medicaid funds from a state if it did not agree to provide for a new class of beneficiaries would constitute the constitutional wrong of coercion if animated or infected by a bad purpose. Taken together, then, the first and second points are these: *compulsion and coercion are not the same things, and the constitutional wrong that conditional spending offers more plausibly instantiate is that of coercion, not of compulsion.*

Third, the basic principles that govern whether a conditional spending offer from the national government to the states is unconstitutionally coercive are not particular to the conditional spending context. Instead, they lie at the heart of a general solution to the ubiquitous puzzle of

---

16. That joint opinion was styled a dissent. But on the particular question on which I am focusing—whether Congress may constitutionally threaten to withhold all Medicaid funding on a state's refusal to accept federal funds to provide Medicaid coverage to a new class of beneficiaries—the votes of these four "dissenters" were necessary to constitute a majority. Accordingly, I will refer to the opinion of Justices Scalia, Kennedy, Thomas, and Alito as "the joint opinion." Given this Article's focus, I reserve the term "dissent" for the opinion by Justice Ginsburg, writing only for herself and Justice Sotomayor on this point.

17. *Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. at 2629 (Ginsburg, J., dissenting in part).

18. *See id.* at 2605–06 (Roberts, C.J.) (discussing Congress's purpose of using the Medicaid expansion to drastically expand coverage and essentially recreate Medicaid).

"unconstitutional conditions"—that is, the puzzle regarding whether and under what circumstances it is constitutionally permissible for government to condition a benefit on an offeree's exercising or not exercising its constitutional rights in some preferred way.[19] Fourth, application of these general "trans-substantive" principles to the instant case suggests that the Medicaid expansion probably was coercive and therefore the Court was probably right—though not for the reasons it gave—to hold that that provision exceeds our best understanding of constitutional limits on Congress's power.[20]

These four theses are developed over five parts. Part I unpacks the arguments advanced in the two opinions that together made up a majority on the Spending Clause question and elucidates the key concepts upon which much of the analyses in the body of the Article will rely—namely, coercion, and compulsion. (Following convention, I will underline these words when I am invoking the concepts and when I think that, given the context, a reminder will be useful.) This Part shows that the majority on this point effectively interpreted what the joint opinion terms "the anti-coercion principle"[21] in Spending Clause jurisprudence as an "anti-compulsion principle"—that is, as a rule that disables Congress from inducing the states to act in accord with the wishes of the national government by offering benefits on terms that the states could not, as a practical matter, reject.

Part II casts doubt on the soundness of such a rule. Contract law, on which the Chief Justice and the joint opinion both rely, does not offer the support they claim. Very likely, the best argument for it is the one advanced by the state challengers to the Act. Without meaningful limits on Congress's spending power, they argued, federalism-based limits on Congress's other powers "would be for naught."[22] Therefore, "a judicially enforceable outer limit on Congress's power to use federal tax dollars to coerce States is . . . a constitutional necessity."[23] There is merit to that argument. But it does not quite support the conclusion drawn. A judicially enforceable limit on

---

19. The heart of a general solution, but not the entirety of it: a conditional offer that does not amount to coercion might be unconstitutional on other grounds. Coercion is the *distinctive*, but not the *sole*, constitutional wrong that conditional offers might instantiate. *See generally* Mitchell N. Berman, *Coercion Without Baselines: Unconstitutional Conditions in Three Dimensions*, 90 GEO. L.J. 1 (2001).

20. This conclusion takes as a given the correctness of the anti-commandeering decisions, *New York v. United States*, 505 U.S. 144 (1992), and *Printz v. United States*, 521 U.S. 898 (1997). I am sympathetic to the suggestion that the best understanding of our constitutional order would leave Congress with more authority to mandate behavior by the states than current case law allows. But this Article analyzes Spending Clause jurisprudence under the assumption that Congress could not mandate state participation in the Medicaid program.

21. The joint opinion deploys this term unhyphenated. I have taken the liberty of inserting a hyphen because doing so makes it easier to distinguish visually the two construals of this principle that I identify.

22. Brief of State Petitioners on Medicaid at 20, *Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. 2566 (No. 11-400).

23. *Id.*

Congress's ability to coerce the states through conditional spending grants need not assume the form of an anti-compulsion rule given the availability of an anti-coercion rule instead. Indeed, as Part II further shows, constitutional doctrines outside the spending context strengthen the appeal of an anti-coercion principle while undermining the plausibility of an anti-compulsion principle. Part III seeks to make readers more receptive to a true anti-coercion principle, and to mitigate objections from the church of stare decisis, by developing the claim, already noted, that the Roberts opinion actually flirts with this alternative construal of the critical principle.

Part IV—the longest and most complex part of the Article—examines whether the conditional offer embodied in the Medicaid expansion constitutes impermissible coercion. Because, as noted above, I believe that the conditional spending problem is, in critical respects, just an instance of the more general problem of "unconstitutional conditions," the first task of this Part is to develop and defend a general account of the circumstances in which it can be unconstitutionally coercive for government to offer "benefits" on condition that the offeree not exercise one of its constitutional rights.[24] That general account centers on a denial of the oft-stated and widely held belief that, if a rightholder (be it an individual or a state) is not entitled to some particular boon, then government may withhold it for any reason at all without offending the Constitution. To the contrary, I argue, government unconstitutionally penalizes the exercise of a right if it withholds a benefit for certain bad purposes or reasons. In other words, I challenge the conventional scholarly wisdom that maintains that the concept of penalty is incoherent or normatively inert. The Part's second task, accordingly, is to apply that general account to the Medicaid expansion. In concluding (tentatively) that the statute runs afoul of general principles regarding coercion and penalty, Part IV, in effect, returns to critics of the Medicaid expansion what Part II had taken away.

Part V considers objections, and articulates refinements, to my general analysis of coercive offers—including the general analysis of penalties—and to the application of that account to the Medicaid expansion.[25] That final Part will underscore a point that warrants emphasis at the outset: I present the analyses that follow not as a watertight argument in support of a single "bottom-line" conclusion, but as a *framework* for analyzing conditional offers by the state—a framework that is filled out more fully and confidently

24. An account could be "general," though not universal or exceptionless. *See infra* Part V, Objection 4.

25. For an early presentation of some of the ideas developed here, see my comments posted to the blog Balkinization while *NFIB* was pending. Mitch Berman, *Coercion, Compulsion, and the ACA*, BALKINIZATION (Apr. 6, 2012, 3:49 AM) http://balkin.blogspot.com/2012/04/coercion-compulsion-and-aca.html; Mitch Berman, *More on Unconstitutional Conditions and the ACA*, BALKINIZATION (Apr. 8, 2012, 10:05 AM) http://balkin.blogspot.com/2012/04/more-on-unconstitutional-conditions-and.html. I am very grateful to Sandy Levinson for prodding me to post on the topic and to Jack Balkin for providing an excellent forum for a productive exchange.

here, sketched more thinly or tentatively there. Readers who end up rejecting my (avowedly uncertain) judgment that the Medicaid expansion was unconstitutionally coercive need not, for that reason alone, reject *in toto* the machinery I propose. The analysis that follows consists of a fair number of moving parts. They do not all stand or fall together.

I. Of Coercion and Compulsion

Those portions of the three opinions that address whether it is constitutional for Congress to threaten to withhold all of a state's Medicaid funding for existing beneficiaries (the blind, the disabled, the elderly, pregnant women, and needy families with dependent children) unless it accepts new funding, with associated conditions, for a new class of beneficiaries (adults, including those without children, with incomes up to 133% of the federal poverty level) are long, totaling over fifty pages together. Despite their combined length, however, one single theme leaps out most plainly: this case seemingly turns, for all the Justices, on a vice they call "coercion." Both the Roberts opinion and the joint opinion squarely conclude both that this particular condition is unconstitutional because it is "coercive" or constitutes impermissible "coercion," and that what makes this so is that it leaves the states with "no real choice" but to accept. Making clear that this is how she reads the majority,[26] Justice Ginsburg objects that "[t]he coercion inquiry . . . appears to involve political judgments that defy judicial calculation."[27]

Accordingly, the first step toward understanding the grounds of, and possible difficulties with, the Court's reasoning in support of its Spending Clause holding must be to get clear on just what the Court means by "coercion."

A. Conceptual and Terminological Preliminaries

Anyone familiar with Supreme Court case law on conditional spending prior to *NFIB* will have noticed this striking feature: the Court routinely uses the terms "coercion" and "compulsion" in a loose fashion, sometimes treating them as synonyms, sometimes not, and never carefully defining either.

Take, to start, the very brief passage from *Dole* in which the Court appears to proscribe conditions "so coercive as to pass the point at which 'pressure turns into compulsion.'"[28] Although this passage is routinely read—including by Chief Justice Roberts and by the authors of the joint opinion—to prohibit "coercion," its literal import is to proscribe "compulsion," the overwhelming implication being that coercive offers that

26. *Nat'l Fed'n of Indep. Bus.*, 132 S.Ct. at 2639–40 & n.24 (Ginsburg, J., dissenting in part).
27. *Id.* at 2641.
28. South Dakota v. Dole, 483 U.S. 203, 211 (1987).

do not amount to compulsion are permissible. That is just a single passage, so should not be over-read were it unusual. In fact, though, the failure carefully to distinguish coercion from compulsion is entirely representative of the case law.[29]

That Supreme Court Spending Clause opinions fail to distinguish between coercion and compulsion in any analytically satisfactory manner is further evidenced by a glance at the work of the best constitutional lawyers. For a striking illustration, consider the principal brief filed by the state challengers in the health care litigation, authored by former Solicitor General Paul Clement. From *Dole*'s declaration that an exercise of Congress's spending power would violate the Constitution if it were "so *coercive* as to pass the point at which 'pressure turns into *compulsion*,'" that brief draws the lesson that "Congress may not use its spending power *coercively*."[30] It also deems "the *coercion* doctrine" violated on the grounds that "the ACA . . . *compels* the States to act in ways that Congress could not *compel* directly."[31] Further examples of the brief's apparent conflation of coercion and compulsion could be multiplied with ease: these few passages are all culled from a single page.[32]

Chief Justice Roberts endorses the very same conflation of coercion and compulsion, or equivocation between them, when stating the issue:

> The States . . . contend that the Medicaid expansion exceeds Congress's authority under the Spending Clause. They claim that Congress is *coercing* the States to adopt the changes it wants by threatening to withhold all of a State's Medicaid grants, unless the State accepts the new expanded funding and complies with the conditions that come with it. This, they argue, violates the basic principle that the "Federal Government may not *compel* the States to enact or administer a federal regulatory program."[33]

This pattern of usage is frequently a strong indication that the speaker or author lacks a firm grasp on the precise idea or concept she is groping for. She has a rough sense of the idea, or knows the vicinity, but hasn't nailed it down. I don't mean this as a biting criticism. It is hard work always to identify the precise concept that we have dimly or loosely in mind, and not

---

29. *See, e.g.*, Coll. Savings Bank v. Fla. Prepaid Postsecondary Educ. Expense Bd., 527 U.S. 666, 687 (1999) (quoting *Dole*'s "point at which 'pressure turns into compulsion'" passage, and then concluding that "the point of *coercion* is automatically passed—and the voluntariness of waiver destroyed—when what is attached to the refusal to waive is the exclusion of the State from otherwise lawful activity" (emphasis added)).

30. Brief of State Petitioners on Medicaid at 27, *Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. 2566 (2012) (No. 11-400) (emphasis added).

31. *Id.* (emphasis added).

32. *Id.*

33. *Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. at 2601 (Roberts, C.J.) (quoting New York v. United States, 505 U.S. 144, at 188 (1992)) (emphases added).

always worth the effort. Not infrequently, the loose grasp is good enough for our purposes.

But not infrequently it isn't. And that is what should worry us here. If the words "coercion" and "compulsion" are not synonymous, but rather capture or are best associated with different concepts, then we cannot tolerate looseness or imprecision in any case in which the two pull apart. When confronting any conditional offer that plausibly coerces the states to accept without compelling acceptance or conversely, any offer that subjects the states to compulsion but not to coercion, it becomes essential to identify which is the constitutional wrong—coercion or compulsion, or perhaps the union of the two, or something else entirely—and then to carefully establish that the features of the program or provision under review make out the concept that is constitutionally significant and not the related concept that might be constitutionally irrelevant.

In the remainder of this section, I aim to establish that coercion and compulsion *are* different concepts. This is a modest claim. To forestall possible misunderstanding, I should emphasize that I am not offering definitions of the *words* "coercion" and "compulsion." I am offering accounts of two distinct concepts to which I am affixing the distinct words "coercion" and "compulsion" as handy labels. Of course, I do believe that the ordinary meanings of the words correspond closely enough to the concepts as I demarcate them to make it reasonable to employ these words and not others. I hope and rather expect that readers will share those judgments. But please keep in mind that our goal here is to focus on the concepts rather than the words. I am trying to make two concepts, and the respects in which they are different, tolerably clear. If you understand the concepts to which I will refer by the words "coercion" and "compulsion," then the argumentative uses to which I will put these concepts will not be jeopardized if you also harbor doubts about the extent to which you would define our existing words "coercion" and "compulsion" to match the concepts as I roughly describe them. (Similarly, although I think I am offering accounts of two distinct concepts, I believe that nothing turns on whether you share that judgment. If you believe that I am misdescribing the concepts that I am calling "coercion" and "compulsion," you may treat the two phenomena that I distinguish as simply that—phenomena. The important questions will turn out to be whether the fact, if true, that a conditional spending offer instantiates this or that phenomenon warrants the judgment that the conditional offer is constitutionally problematic. What are the best accounts of the concepts of coercion or compulsion should not distract us.)

Coercion is generally thought to be a type of wrong. It's something that we presumptively ought not to engage in, and that properly subjects us to

criticism, censure or, at a minimum, a demand for justification, if we do.[34] Of course, there are many and diverse types of wrongs. To a first approximation, coercion is the wrong of exerting *wrongful* pressure on a subject to do as the coercer wishes.[35] And the usual way in which one puts wrongful pressure on a target's choices is by threatening to wrong him if he does not comply with the threatener's "demand" or "condition."[36] Roughly, then, a threat is coercive, or constitutes coercion, if it would be wrongful for the threatener to carry it out.[37]

Compulsion, in contrast, is not a wrong—at least not all by itself. It is a description, if possibly a normatively freighted one, of certain circumstances of action, namely those in which, for one reason or another, our choices are very substantially constrained.[38] Again to a first approximation, one is compelled to do such-and-such, or is subject to compulsion, when there is some coherent sense in which one could not have done otherwise. Compulsion can be produced in various ways. For example, it can be the product of extremely powerful irrational urges, like those arising from addiction or other forms of mental disorder.[39] Alternatively, it can be the product of rational pressure to pursue the course of action that powerfully dominates all alternatives in a severely circumscribed choice set.[40] Depending on other factors, the descriptive fact that one has acted in the face of compulsion may or may not serve, normatively, to make out a type of excusatory or mitigating condition.[41] In short, compulsion is a state of affairs to which, ideally, we would not be subject, and that, when present, can potentially ground relief from responsibility or liability.

Again, these are first-pass accounts of the two concepts. Either or both might benefit from refinement. For our purposes, though, exquisite precision

---

34. Mitchell N. Berman, *The Normative Functions of Coercion Claims,* 8 LEGAL THEORY 45, 47 (2002).

35. *Id.*

36. *See* Martin Gunderson, *Threats and Coercion*, 9 CAN. J. PHIL. 247, 248 (1979) (describing dispositional coercion as involving "the threat of sanctions").

37. This is the dominant understanding in the philosophical literature. For overviews, see Scott Anderson, *Coercion*, *in* STANFORD ENCYCLOPEDIA OF PHILOSOPHY (2011), *available at* http://plato.stanford.edu/archives/win2011/entries/coercion; William A. Edmundson, *Coercion*, *in* THE ROUTLEDGE COMPANION TO PHILOSOPHY OF LAW 451 (Andrei Marmour ed., 2012). Important works that defend and develop this claim include ALAN WERTHEIMER, COERCION (1987); Gunderson, *supra* note 36; and Vinit Haksar, *Coercive Proposals [Rawls and Gandhi]*, 4 POL. THEORY 65, 68–70 (1976). For my contribution to the general philosophical literature, see generally Berman, *supra* note 34.

38. *See* Robert Audi, *Moral Responsibility, Freedom, and Compulsion*, 11 AM. PHIL. Q. 1, 3 (1974) (suggesting that people who act with limited choices may be acting with less freedom); *see also* Vincent Brümmer, *On Not Confusing Necessity with Compulsion: A Reply to Paul Helm*, 31 RELIGIOUS STUD. 105, 105–06 (1995) (suggesting that choice can be limited by factual circumstances without destroying freedom of choice).

39. *See* Audi, *supra* note 38, at 5 (illustrating that internal compulsions such as obsessions, phobias, and irresistible impulses can lead to unavoidable actions).

40. Matt Zwolinski, *Sweatshops, Choice, and Exploitation*, 17 BUS. ETHICS Q. 689, 701 (2007).

41. *Id.*

is not essential. These provisional accounts are sufficient to establish the critical point that these are distinct concepts. And that claim is demonstrated by the fact that there exists both compulsion-without-coercion and coercion-without-compulsion.

Here's just one quick example of compulsion-without-coercion. Law student, L, accepts a job with a firm that represents clients to whom L strenuously objects or that, in any other fashion, runs contrary to important principles or values of L's. L wouldn't accept the job but for the facts that it is L's only offer and that L has very substantial loan obligations. L can properly answer, in response to the charge that she has compromised her principles, that she "was compelled" to do so or "had no choice." Nonetheless, L wasn't "coerced into" accepting the job and nobody—not the firm or anybody else—is properly charged with coercion.

And here's an example of coercion-without-compulsion: T, a thug, threatens H with some moderate violence—say, a broken finger—unless H turns over his briefcase. H complies. Unbeknownst to T, the briefcase contains most of H's and W's savings. When H returns home and reports the robbery, H's spouse, W, is aghast. "How could you possibly have given up all our funds for Junior's education?!" W demands. If H responds that he "was compelled to do so" or "had no choice," W could be right (depending upon the details, of course) to reject the claim. H was not compelled to give up that money. Given the threat he faced, H should have run or resisted. Yet T did engage in coercion. T didn't merely try to coerce H, for he did, after all, succeed. Assuming that T threatened H with unpleasantness that T was wrong to threaten but that H could have endured and should have under the circumstances, T coerced H into giving up his money, though H wasn't compelled to do so.[42]

Naturally, countless interactions amount to both coercion and compulsion—what we might term either coercion-through-compulsion or compulsion-by-coercion. "Your money or your life" is a paradigm. That is to be expected because coercive proposals are intended to induce compliance with a condition or demand, and the issuer of the proposal—the coercer—understands that success in this aim is a function of the pressure that the target of the coercion experiences, and not the bare wrongness of the consequence threatened.[43] But the key point is that coercion and compulsion are analytically distinct and can and do come apart in the real world.

---

42. In response to Justice Ginsburg's observation that it would cost states little to accept the Medicaid expansion, Chief Justice Roberts objected that "the size of the new financial burden imposed on a State is irrelevant in analyzing whether the State has been coerced into accepting that burden. 'Your money or your life' is a coercive proposition, whether you have a single dollar in your pocket or $500." Nat'l Fed'n of Indep. Bus. v. Sebelius, 132 S. Ct. 2566, 2605 n.12 (2012) (Roberts, C.J.). He is quite right. My point here is that "Your money or I'll break your arm" is also "a coercive proposition." But depending upon the context it might not be one that amounts to compulsion.

43. Gunderson, *supra* note 36, at 253–54.

Coercion and compulsion are both characterizations of features of events in which one agent exerts pressure on another to do as the first agent wishes. At the risk of some simplification, compulsion is constituted by the *amount* of pressure and coercion is constituted by its *character*.

The critical question, therefore, is this: In the context of constitutional challenges to the Medicaid provisions of the ACA (and in the spending context more generally, and—just possibly—in other conditional offer contexts more generally still), which is or should be the operative concept—coercion or compulsion? This question cannot be answered by simply pointing out that "it's called 'the anti-coercion principle,' stupid." As we will see, the word "coercion" is sufficiently plastic or ambiguous to encompass both concepts, coercion and compulsion (and perhaps other concepts as well).

B. The "Anti-Coercion Principle" as an Anti-Compulsion Principle

Given the ambiguity of the word "coercion," the joint opinion starts, very helpfully, by expressly acknowledging that "coercion" requires definition. "Once it is recognized that spending-power legislation cannot coerce state participation," the opinion observes, "two questions remain: (1) What is the meaning of coercion in this context? (2) Is the ACA's expanded Medicaid coverage coercive?"[44] Without missing a beat, it then announces that

> The answer to the first of these questions—the meaning of coercion in the present context—is straightforward. As we have explained, the legitimacy of attaching conditions to federal grants to the States depends on the voluntariness of the States' choice to accept or decline the offered package. Therefore, if States really have no choice other than to accept the package, the offer is coercive, and the conditions cannot be sustained under the spending power.[45]

In short, despite its reference to the "anti-coercion principle," the standard the joint opinion actually deploys would be more accurately termed (in the language of this Article) an "anti-compulsion principle." Justice Ginsburg is not far off when observing that, "[f]or the joint dissenters, . . . all that matters, it appears, is whether States can resist the temptation of a given federal grant."[46]

Furthermore, the Chief Justice's opinion, for himself and Justices Breyer and Kagan on this point, seems largely in accord. "Permitting the Federal Government to force the States to implement a federal program

---

44. *Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. at 2661 (joint opinion).

45. *Id.*

46. *Id.* at 2640 n.24 (Ginsburg, J., dissenting in part).

would threaten the political accountability key to our federal system," it reasons.[47]

> "[W]here the Federal Government directs the States to regulate, it may be state officials who will bear the brunt of public disapproval, while the federal officials who devised the regulatory program may remain insulated from the electoral ramifications of their decision." Spending Clause programs do not pose this danger when a State has a legitimate choice whether to accept the federal conditions in exchange for federal funds. . . . But when the State has no choice, the Federal Government can achieve its objectives without accountability . . . .[48]

And in this case, the opinion concludes, the states really do lack a choice. "The threatened loss of over 10 percent of a State's overall budget . . . is economic dragooning that leaves the States with no real option but to acquiesce in the Medicaid expansion."[49] By striking down this condition, the opinion thus "limits the financial pressure the [federal government] may apply to induce States to accept the terms of the Medicaid expansion."[50] Just like the authors of the joint opinion, then, the Chief Justice understands the anti-coercion principle from conditional spending jurisprudence to police compulsion.

II. Compulsion, Really?

Suppose the states have "no choice" but to agree to provide coverage for the ACA's new class of Medicaid beneficiaries because the cost to them of doing without Medicaid funds at all is so enormous, and therefore that the Medicaid expansion subjects them to compulsion. Of course, the states do not *literally* have no choice in the matter. But if compulsion exists only when an offeree has "no choice" but to accept, and if "no choice" in this context means, well, *no* choice, then compulsion would be a nearly useless concept. Even seemingly paradigmatic instances of compulsion (including "your money or your life") would turn out not to be compulsion at all. And certainly the states could never be compelled by the threat of a withdrawal of federal funds, contrary to the assumption in *Dole* and *Steward* that this is a theoretical possibility. The lesson is that "no choice" must be taken idiomatically, not literally, and be given a looser construction. Thus, compulsion exists when an offeree has no *reasonable* choice or no choice that it would be remotely rational for it to adopt, or something like this.[51]

---

47. *Id.* at 2602 (Roberts, C.J.).

48. *Id.* at 2602–03 (internal citations omitted) (quoting New York v. United States, 505 U.S. 144, 169 (1992)).

49. *Id.* at 2605.

50. *Id.* at 2608.

51. Recognizing that "no choice" cannot be taken literally, the joint opinion and Chief Justice Roberts sometimes qualify the phrase; "no real choice" is a favorite alternative. *See, e.g.*, *id.* Insofar as "real" contrasts with "fake," it cannot be the most apt qualifier to have been selected. But it does adequately signal that there are difficulties here that require attention. For analysis of

However the "no choice" standard is interpreted, it will be sufficiently vague as to license doubts that it meets the "judicial manageability" bar for judicially enforced constitutional doctrine.[52] But put that worry aside. So long as an anti-coercion principle remains part of judicial Spending Clause doctrine, and if it forbids compulsion, then whatever the difficulty of evaluating borderline cases, it is hard to contest the majority's conclusion on the facts of this case. The Medicaid expansion threatened states with the aggregate loss of $233 billion per year, equaling over 10% of all state budgetary outlays.[53] The judgment that it would be so damaging for a state to sustain the loss of so many funds as to compel it to accept the new deal, if not quite inescapable,[54] is more than reasonable. If the majority holding is wrong, then, it is more likely because the majority was wrong to conclude that Congress is barred from making offers that the states are compelled to accept, without more. The question is this: *why* should we understand the anti-coercion principle as one that disables Congress from using its spending power to craft offers so attractive that states are compelled to accept?

In posing the question this way, I do not mean to gain any mileage from characterizing the proposal as an "offer" rather than as a "threat." I prefer to adopt the convention according to which, strictly speaking, every biconditional proposal consists of both a conditional offer and a conditional threat: the offer (threat) is the conditional proposal that contains the consequent that the proposal-maker anticipates the recipient will find the more (less) attractive of the two. Thus, the merchant's "two-for-one" offer is also a threat not to give you two if you don't buy one; the robber's threat to kill you if you don't hand over your money is also an offer to let you live if you do. Of course, it would ring false to describe the first proposal as a "threat" or the latter as an "offer." But I think the much-explored question of whether a particular proposal *as a whole* is better characterized as a threat or an offer distracts us from the normatively important questions.[55]

---

different ways to cash out the "no choice" standard, and of difficulties that attend to each, see Lynn A. Baker & Mitchell N. Berman, *Getting off the* Dole*: Why the Court Should Abandon Its Spending Doctrine, and How a Too-Clever Congress Could Provoke it to Do So*, 78 IND. L.J. 459, 517–21 (2003).

52. *See Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. at 2641 (Ginsburg, J., dissenting in part) ("The coercion inquiry . . . appears to involve political judgments that defy judicial calculation.").

53. *Id.* at 2605 (Roberts, C.J.); *id.* at 2664 (joint opinion). Although the joint opinion describes this sum as "equaling 21.86% of all state expenditures combined," that figure reflects the percentage of state spending that is comprised by state and federal Medicaid funds aggregated. Brief of State Petitioners on Medicaid at 15, *Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. 2566 (No. 11-400).

54. For an intriguing presentation of doubts, see Brian Galle, *Does Federal Spending "Coerce" States?: Evidence from State Budgets*, 107 NW. U. L. REV. (forthcoming 2013).

55. For elaboration and defense of this position, see Berman, *supra* note 34, at 55–59. *See also* E. Allan Farnsworth, *Coercion in Contract Law*, 5 U. ARK. LITTLE ROCK L.J. 329, 333 (1982) ("Nothing is gained by attempting to distinguish offers from threats for the purposes of the law of duress. Since a claim of duress can only succeed if the threat was one that the law condemns, the significant task is not to distinguish offers from threats but to distinguish those threats that the law condemns from those that it does not condemn."). The canonical effort to distinguish *threats* from

Accordingly, we can rephrase the question: why should we understand the anti-coercion principle to disable Congress from using its spending power to threaten states with consequences so unattractive that they are compelled to comply with the stated condition?

A. "... Much in the Nature of a Contract"

In seeking an answer, we might start with the joint opinion. Recall its assertion that

> The answer to the first of these questions—the meaning of coercion in the present context—is straightforward.... [T]he legitimacy of attaching conditions to federal grants to the States depends on the voluntariness of the States' choice to accept or decline the offered package. Therefore, if States really have no choice other than to accept the package, the offer is coercive, and the conditions cannot be sustained under the spending power.[56]

I observed that, in this short passage, the opinion contends that "coercion" means compulsion. Indeed, it claims that this is straightforward or uncontroversially true. What, we might now ask, makes this correct, let alone straightforwardly so?

In large measure, the joint opinion's answer is: the Court's conditional spending precedent, *Dole* in particular.[57] But *Dole* is a slender reed on which to rest. We have already seen that *Dole*, like other spending cases, used the words "coercion" and "compulsion" so cavalierly as to instill significant doubt that the authors knew precisely what concepts they were after.[58] Moreover, the somewhat ambivalent manner in which *Dole* invoked the anti-coercion principle (however that principle may be construed) provides further reason not to put all of one's pineapples in this particular basket. It would have been easy enough for the *Dole* majority to plainly announce *five* requirements that any condition attached to federal spending grants to the states must satisfy: it must promote the general welfare, be unambiguous, be germane to the federal interest in the spending program, not induce the states to violate the Constitution, and not coerce the states into accepting. Instead, Chief Justice Rehnquist's opinion listed the first four restrictions in a single paragraph and then, only after determining that none condemned the condition on highway funds at issue in that case, introduced *Steward Machine*'s ruminations on coercion almost as an afterthought.[59] Justice Ginsburg draws from this expositional curiosity the conclusion that *Dole*

---

offers is Robert Nozick, *Coercion*, in PHILOSOPHY, SCIENCE, AND METHOD 440, 447–53 (Sidney Morgenbesser et al. eds., 1969).

56. *Nat'l Fed'n of Indep. Bus.*, 162 S. Ct. at 2661 (joint opinion).

57. *Id.* at 2659, 2661 (citing South Dakota v. Dole, 483 U.S. 203, 211–12 (1987)).

58. *See supra* notes 28–29 and accompanying text.

59. *Dole*, 483 U.S. at 207–11.

only "mentioned, but did not adopt, [this] further limitation."[60] That might be too grudging. But a weaker and more defensible lesson is that, if alternative interpretations of the anti-coercion principle are reasonably available, *Dole* alone provides less robust support for the interpretation adopted than one would hope for.

Happily, and to its credit, the joint opinion does not rest its interpretation solely on passages from Spending Clause precedent that could conceivably be characterized as dicta. Instead, it invokes contract law principles. "When federal legislation gives the States a real choice whether to accept or decline a federal aid package," it explains, "the federal-state relationship is in the nature of a contractual relationship. . . . And just as a contract is voidable if coerced, the legitimacy of Congress's power to legislate under the spending power . . . rests on whether the State *voluntarily* and knowingly accepts the terms of the 'contract.'"[61]

Parsed as an argument, the joint opinion's reasoning on this score runs something like this: (1) Congress's power to legislate under the spending power is informed by contract law principles; (2) contract law prohibits coercion; (3) therefore, rules governing exercises of the spending power properly prohibit coercion; (4) the meaning of coercion for purposes of contract law is compulsion; (5) therefore, the meaning of coercion in the spending context is compulsion.

Premise (4), though unstated, is implicit. After all, by observing that meaning must be expressly ascribed to "coercion" in the spending context, the joint opinion acknowledges that the term is ambiguous or at least not transparent. It also says—or, at a minimum, strongly implies—that the limits on Congress's spending power arise from principles of contract law, or from the same more fundamental considerations that undergird contract law: "[J]ust as a contract is voidable if coerced, the legitimacy of Congress's power to legislate under the spending power . . . rests on whether the State *voluntarily* and knowingly accepts the terms of the 'contract.'"[62] So premise (4) is necessary support for (5).

But premise (4) is false. Contract law does recognize a defense termed, interchangeably, "coercion" or "duress." As the Restatement of Contracts provides, "If a party's manifestation of assent is induced by an improper threat by the other party that leaves the victim no reasonable alternative, the contract is voidable."[63] What makes a threat "improper" is notoriously fuzzy. A threat to commit a crime or tort would count, of course, but so too would a "breach of the duty of good faith and fair dealing" under an existing contract, and, when it produces unfair terms, a threat to perform an act that

60. *Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. 2566, 2634 (2012) (Ginsburg, J., dissenting in part).

61. *Id.* at 2659–60 (joint opinion) (emphasis in original) (citations and internal quotation marks omitted). Roberts relies squarely on the contract law analogy too. *See id.* at 2602 (Roberts, C.J.).

62. *Id.* at 2660 (internal quotation marks omitted).

63. RESTATEMENT (SECOND) OF CONTRACTS § 175(1) (1981).

"would harm the recipient and would not significantly benefit the party making the threat."[64] The important point, though, is that the fact that one party had "no choice" but to accept a contract or a contractual condition is never sufficient alone to make the contract voidable.[65] There must always be, in addition to the lack of "reasonable alternative[s]," an "improper threat."[66] In short, duress or coercion, in contract law, requires something very much like the conjunction of coercion and compulsion.

The doctrine of unconscionability likewise will not support the idea that legal consequences should follow from the mere fact that one party to an agreement has "no choice" other than to accept.[67] Comment 1 to § 2-302 of the Uniform Commercial Code offers an essentially circular definition: "The basic test is whether, in the light of the general commercial background and the commercial needs of the particular trade or case, the clauses involved are so one-sided as to be unconscionable under the circumstances existing at the time of the making of the contract."[68] Farnsworth's treatise states that

> [t]he most durable answer [for what unconscionability is] is probably that of the court in *Williams v. Walker-Thomas*: "Unconscionability has generally been recognized to include an absence of meaningful choice on the part of one of the parties [a.k.a. procedural unconscionability] together with contract terms which are unreasonably favorable to the other party [a.k.a. substantive unconscionability]."[69]

Most significantly for present purposes, "judges have been cautious in applying the doctrine of unconscionability, recognizing that the parties often must make their contract quickly, that their bargaining power will rarely be equal, and that courts are ill-equipped to deal with problems of unequal distribution of wealth."[70] In particular, "[c]ourts have resisted applying the doctrine [of unconscionability] where there is only procedural unconscionability without substantive unfairness."[71]

Both the joint opinion and Roberts's opinion place great weight on the Court's much-quoted observation in *Pennhurst State School & Hospital v. Halderman*[72] that "legislation enacted pursuant to the spending power is much in the nature of a contract: in return for federal funds, the States agree to comply with federally imposed conditions."[73] From this premise, the

---

64. *Id.* § 176(1)(d), (2)(a).

65. *Id.* at § 2-302 cmts. a–b.

66. *Id.*

67. I am grateful to John Golden for encouraging me to emphasize this point.

68. U.C.C. § 2-302 cmt. 1 (1996).

69. E. ALLAN FARNSWORTH, CONTRACTS § 4.28, at 301 (4th ed. 2004); *cf. id.* at 299 (describing unconscionability as "incapable of precise definition").

70. *Id.* at 302.

71. *Id.*

72. 451 U.S. 1 (1981).

73. *Id.* at 17.

Pennhurst Court concluded that "[t]he legitimacy of Congress's power to legislate under the spending power thus rests on whether the State voluntarily and knowingly accepts the terms of the 'contract.'"[74] Plucking the adverb "voluntarily" from its contract law context, the *NFIB* joint opinion concludes that an exercise of the spending power is unconstitutional if the offeree has "no choice" but to accept.[75] Contract law principles do not support that expansive reading of what makes acceptance *in*voluntary. A contract is not voluntary for purposes of contract law if it is the product of duress or unconscionability. And both doctrines require some form of impropriety by the offeror—an impropriety that is not made out just by the fact that the offeror crafted terms that it knew the offeree could not reasonably reject.[76]

The bottom line is that "coercion" in contract law does not mean compulsion, and there is no principle of contract law that permits a contract to be voided just because one party had "no choice" but to accept. This being so, the joint opinion is not entitled to its blithe assertion that the "anti-coercion principle" is offended by an offer that effectively "compels" acceptance or, put otherwise, that "coercion" in Spending Clause jurisprudence means compulsion. It might. But analogizing a state's agreement to comply with conditions on the receipt of federal funds to private agreements governed by contract law furnishes no support for this assertion. To the contrary, if it is true, as the joint opinion suggests, that the limitations on Congress's spending power derive from the same source as do the limits on "coerced" contracts, and if it is true, as the Restatement provides, that "coercion" in contract law requires coercion, then the conclusion to draw is radically opposed to that which the joint opinion asserts: "coercion" in Spending Clause jurisprudence requires coercion, and not compulsion (or not only compulsion).[77]

Again, all that I have just written still falls short of conclusively establishing that a majority in *NFIB* was wrong to enforce an anti-

---

74. *Id.*

75. Nat'l Fed'n of Indep. Bus. v. Sebelius, 132 S. Ct. 2566, 2661 (2012) (joint opinion).

76. The *Pennhurst* Court might have appreciated all this. For after briefly referencing voluntariness, the Court's opinion says nothing more about it and proceeds to examine knowingness, ultimately dismissing a lawsuit against a state defendant on the grounds that the particular duties that plaintiffs alleged the state had assumed when accepting federal funds for the developmentally disabled had not been stated with sufficient clarity. Despite repeated citations to *Pennhurst* both by Roberts and by the authors of the joint opinion, the holding of that case adds essentially nothing to the requirement, subsequently set forth in *Dole*, that conditions on spending be unambiguous.

77. Remarkably, the contract law-inspired case for a compulsion-based interpretation of the spending doctrine's "anti-coercion principle" is weaker still. Even when coercion has been made out in contract law, the remedy is that the contract is *voidable*. Here, the majority substantially weakens the notion of coercion—from, roughly, the conjunction of coercion and compulsion to (mere) compulsion—and also substantially strengthens the remedy. If the joint opinion were serious about the contract analogy, the lesson would be that states could, without adverse consequence, back out of deals to which they had agreed under compulsion. The majority goes beyond that to disable Congress even from making offers that subject the states to compulsion.

compulsion principle against Congress's use of its spending power. It could be that contract law furnishes a much less appropriate analogy for conditional offers of federal funds to the states than the majority assumes. However close or distant the analogy, we should nonetheless pause to reflect on why contract law does not permit any legal consequences to follow from the mere fact that the offeree of a contract proposal had "no choice" but to accept. It takes that approach because a contrary rule would be absurd. People often accept deals because they have no good choice in the matter. Consider law school graduate, L, in my hypothetical above. It would be crazy to prevent the law firm from making an offer just because it would give L "no choice" but to accept. Such a rule would make it nearly impossible to employ persons with radically limited options. Not surprisingly, then, courts adjudicating contract disputes have rejected a bare anti-compulsion principle time and again.[78]

B. Beyond Contract Law

But perhaps the cases are distinguishable based on the source of the pressure. In the law firm case, even if we rightly say that L had "no choice" other than to accept the firm's offer and thus "was compelled" to accept it, we would not rightly say that *the law firm* compelled L to accept. We would say, instead, that financial straits compelled L's acceptance. With respect to the Medicaid expansion, in contrast, defenders of the majority's reasoning might say that the states were not compelled simply by circumstances to accept, but also that the statute, or Congress, compelled them to accept. The pressure was exerted by Congress and not by other forces or circumstances.

This is a tendentious description of the facts of the case. It seems more accurate to say that the states, much like L, would have been compelled to accept by facts about the world. Each state has many citizens and residents who are unable to provide for their own medical care; the state's populace demands that it ensure that health care be made available for these needy folks; and the resulting financial obligations are too great for the state comfortably to handle.[79] Sure, each state would have greater capacity to provide medical care for its needy if the national government did not tax its citizens to fund the national Medicaid program. On the other hand, if Congress didn't create Medicaid, the states might well find themselves back in a race to the bottom, the logic of which would also frustrate their ability to furnish substantial medical assistance to the poor and disabled. So

78. For a representative decision, but explained with Judge Posner's characteristic lucidity, see Selmer Co. v. Blakeslee-Midwest Co., 704 F.2d 924, 926–29 (7th Cir. 1983).

79. *See* Diane Rowland & Adele Shartzer, *America's Uninsured: The Statistics and Back Story*, 36 J.L. MED. & ETHICS 618, 619, 626 (2008) (outlining the growing number of uninsured and noting public opinion being generally in favor of covering those uninsured); KAISER COMM'N ON MEDICAID & THE UNINSURED, HENRY J. KAISER FAMILY FOUND., THE UNINSURED: A PRIMER 14 fig.13 (2010), *available at* http://www.kff.org/uninsured/upload/7451-06.pdf (showing that states pay for 30% of uncompensated care for the uninsured totaling $17.2 billion).

Congress's net contribution to the pressures that combine to give states "no realistic choice" other than to accept the deals proposed in the ACA is highly uncertain.

Furthermore, even granting that Congress played some causal role in contributing to the circumstances that conspire to compel the states to agree to the new Medicaid conditions contained in the ACA, much more still needs to be said to justify the conclusion that Congress should be disabled from making the offer. In other contexts, the fact that one party is causally responsible for pressure exerted upon another is still insufficient, absent coercion, to disable it from exerting pressure that effectively compels another party to accept its offers.[80] Individuals and governments alike are often permitted to be agents of compulsion.

Plea bargaining presents perhaps the best example. Given a sufficiently large differential between the sentence that a defendant would face if convicted after trial and the sentence he is offered to plead guilty, along with a sufficiently high expected probability of conviction if he goes to trial, any given defendant could find it simply irrational to reject the deal. That is, having no other reasonable or rational choice, he would be compelled to accept. Many academic commentators have concluded, on this basis, that plea bargaining is unconstitutionally coercive.[81] In our terminology, however, all that this establishes is that plea bargaining can constitute compulsion. Yet the fact that the pressure that might compel a defendant to accept is exerted by the government, and not merely by the world at large, does not furnish a credible basis for challenging the plea offer.[82]

Don't misunderstand: plea bargaining should not be immune from constitutional scrutiny. Sometimes, even often, it might constitute the wrong of coercion.[83] My claim here is only that the fact, without more, that the threat of a stiff sanction might give a particular defendant no reasonable choice other than to accept is not a plausible basis for invalidating the offer of a much-reduced sanction in exchange for a guilty plea. Courts have appropriately recognized as much. As a unanimous Supreme Court explained over forty years ago:

---

80. *See, e.g.*, United States v. Mezzanatto, 513 U.S. 196, 209–10 (1995) (asserting that the government is permitted to "exert[] pressure on defendants to plead guilty and to abandon a series of fundamental rights" in the absence of "fraud or coercion").

81. *See, e.g.*, Daniel P. Blank, *Plea Bargain Waivers Reconsidered: A Legal Pragmatist's Guide to Loss, Abandonment, and Alienation*, 68 FORDHAM L. REV. 2011, 2016 (2000); John H. Langbein, *Torture and Plea Bargainingi* 46 U. CHI. L. REV. 3, 12 (1978).

82. *See, e.g.*, Bordenkircher v. Hayes, 434 U.S. 357, 364 (1978) (rejecting a prisoner's constitutional challenge to a plea bargain and stating that "by tolerating and encouraging the negotiation of pleas, this Court has necessarily accepted as constitutionally legitimate the simple reality that the prosecutor's interest at the bargaining table is to persuade the defendant to forgo his right to plead not guilty").

83. For that different argument, see Berman, *supra* note 19, at 98–103.

> The State to some degree encourages pleas of guilty at every important step in the criminal process. For some people, . . . apprehension and charge, both threatening acts by the Government, jar them into admitting their guilt. In still other cases, the post-indictment accumulation of evidence may convince the defendant and his counsel that a trial is not worth the agony and expense to the defendant and his family. *All these pleas of guilty are valid in spite of the State's responsibility for some of the factors motivating the pleas*; the pleas are no more improperly compelled than is the decision by a defendant at the close of the State's evidence at trial that he must take the stand or face certain conviction.[84]

The jurisprudence of plea bargaining, then, supports and strengthens the lesson that contract law teaches: ordinarily, the fact that one party effectively compels another party to accept a deal by offering a benefit on terms that the latter could not reasonably reject is not adequate grounds for bringing adverse legal consequences to bear on the offeror—even when the offeror has played a part in making the threatened state of affairs as unattractive to the offeree as it is.

But ordinarily is not invariably. In at least one context other than conditional spending the Supreme Court has endorsed an anti-compulsion principle: the Establishment Clause. If that principle is sound in that context, perhaps it is sound in the conditional federal spending context too.

The key Establishment Clause case, of course, is *Lee v. Weisman*,[85] a 5–4 decision authored by Justice Kennedy. Deeming it "beyond dispute that, at a minimum, the Constitution guarantees that government may not coerce anyone to support or participate in religion or its exercise,"[86] the Court proceeded to hold unconstitutional officially led prayers at high school graduation ceremonies on the grounds that "the government may no more use social pressure to enforce orthodoxy than it may use more direct means."[87] The objectionable social pressure, the Court explained, consisted of "public pressure, as well as peer pressure" exerted "on attending students to stand as a group or, at least, maintain respectful silence during the invocation and benediction."[88] Moving seamlessly between "coercion" and "compulsion," the Court further emphasized that

> [t]his pressure, though subtle and indirect, can be as real as any overt compulsion. . . . [F]or the dissenter of high school age, who has a reasonable perception that she is being forced by the State to pray in a manner her conscience will not allow, the injury is no less real. . . . It is of little comfort to a dissenter . . . to be told that for her the act of

84. Brady v. United States, 397 U.S. 742, 750 (1970) (emphasis added).
85. 505 U.S. 577 (1992).
86. *Id.* at 587.
87. *Id.* at 594.
88. *Id.* at 593.

> standing or remaining in silence signifies mere respect, rather than participation. What matters is that, given our social conventions, a reasonable dissenter in this milieu could believe that the group exercise signified her own participation or approval of it.[89]

Finally, the majority dismissed impatiently the state's contention that "attendance at graduation and promotional ceremonies is voluntary."[90] This argument, it announced,

> lacks all persuasion. Law reaches past formalism. And to say a teenage student has a real choice not to attend her high school graduation is formalistic in the extreme. . . . Attendance may not be required by official decree, yet it is apparent that a student is not free to absent herself from the graduation exercise in any real sense of the term "voluntary," for absence would require forfeiture of those intangible benefits which have motivated the student through youth and all her high school years.[91]

In several respects *Lee* might appear to be a useful precedent for the *NFIB* majority on the spending issue. First, under the *label* "coercion," *Lee* deployed the *concept* of compulsion. Second, the Court rejected a nominal or formalistic approach to the question of whether a right holder enjoyed a meaningful choice in favor of an inquiry into practical realities. Third, having reasoned that a right holder's nominal choice was not voluntary "in any real sense," it concluded that the challenged practice amounted to unconstitutional "coercion" or "compulsion."[92]

Yet the authors of the joint opinion cannot easily avail themselves of the support that *Lee* might offer, for two of them—Justices Scalia and Thomas—have denounced the *Lee* analysis in just the respects that matter here. Indeed, the central thrust of Scalia's opinion for the four *Lee* dissenters was precisely that *the majority deployed an indefensible conception of coercion.*[93] Although he agreed with "the Court's general proposition that the Establishment Clause 'guarantees that government may not coerce anyone to support or participate in religion or its exercise,'" Scalia could "see no warrant for expanding the concept of coercion beyond acts backed by threat of penalty . . . a brand of coercion that, happily, is readily discernible to those of us who have made a career of reading the disciples of Blackstone rather than of Freud."[94] Importantly, Scalia's objection was not that, while the majority properly understood "coercion" to exist when the government exerts too much pressure on a target, it erred in finding the line between tolerable

89. *Id.*

90. *Id.* at 594.

91. *Id.* at 594–95.

92. *Id.* at 599.

93. *Id.* at 632 (Scalia, J., dissenting) (attacking a "boundless, and boundlessly manipulable, test of psychological coercion . . .").

94. *Id.* at 642.

and excessive pressure crossed on the facts of the case. Rather, as he emphasized some years later, his disagreement with the *Lee* majority concerned "the form that coercion must take."[95] And for Scalia, to repeat, the form that coercion must take is a threat to impose a legal penalty.

Not only for Scalia is this the case. As Justice Thomas reiterated a dozen years after *Lee*, in his *Elk Grove Unified School District v. Newdow* concurrence, "*Lee* adopted an expansive definition of 'coercion' that cannot be defended"[96]—"a notion of 'coercion' that . . . has no basis in law or reason."[97] The legally significant kind of coercion (at least for purposes of the Religion Clauses), Thomas insisted, was precisely the kind that Scalia had previously identified: "that accomplished '*by force of law and threat of penalty.*'"[98] Naturally, precisely what this means turns on what Justices Scalia and Thomas mean by "penalty." We'll explore that question in Part IV. For the present, we can conclude merely that *Lee*'s compulsion-prohibiting spin on the Establishment Clause's own "anti-coercion principle" should not be welcome support for the authors of the *NFIB* joint opinion.[99]

C. *Blurring the Lines of Political Accountability*

The previous section showed that analogies to contract law, plea bargaining, and the law of religion do not support the proposition that conditional offers of federal funds—or, equivalently, conditional threats to withdraw or not to provide federal funds—are normatively problematic just because they give states no reasonable choice but to accept.[100] In fact, those analogies do more to undermine the claim. It remains, then, to consider whether there are good arguments for an anti-compulsion rule in the conditional spending context that are particular to that context and do not depend upon principles or considerations that sweep more broadly. Perhaps even if an anti-compulsion rule makes little sense in most or all other legal domains, the relationship between the national government and the states is sui generis in respects that justify such a rule here.

The majority does advance such an argument, one grounded in the theory, first floated in the anti-commandeering decision *New York v. United*

---

95. McCreary Cnty. v. ACLU, 545 U.S. 844, 908–09 (2005) (Scalia, J., dissenting).

96. 542 U.S. 1, 45 (2004) (Thomas, J., concurring).

97. *Id.* at 49.

98. *Id.* (quoting *Lee*, 505 U.S. at 640 (Scalia, J., dissenting)).

99. Well, not for all of them. Justice Kennedy, one of the authors of the *NFIB* joint opinion, was also the author of *Lee*. According to the Blackmun papers, though, and for whatever it may be worth, Kennedy was a late convert to the view he eventually penned. Having been assigned after conference to write the majority opinion upholding the prayers, Kennedy concluded after several months that his "draft looked quite wrong," causing him to switch his vote and thus produce a new 5–4 majority going the other way. Linda Greenhouse, *Documents Reveal the Evolution of a Justice*, N.Y. TIMES, March 4, 2004, http://www.nytimes.com/2004/03/04/us/documents-reveal-the-evolution-of-a-justice.html?pagewanted=all&src=pm.

100. For a discussion of these analogies, see *supra* notes 78–84 and accompanying text.

States,[101] that national coercion of the states "blurs the lines of political accountability."[102] We already saw that Chief Justice Roberts relies on this consideration.[103] So too does the joint opinion. Quoting *New York* extensively, the joint opinion reasons that

> Where all Congress has done is to "encourag[e] state regulation rather than compe[l] it, state governments remain responsive to the local electorate's preferences; state officials remain accountable to the people. [But] where the Federal Government compels States to regulate, the accountability of both state and federal officials is diminished.". . . When Congress compels the States to do its bidding, it blurs the lines of political accountability. If the Federal Government makes a controversial decision while acting on its own, "it is the Federal Government that makes the decision in full view of the public, and it will be federal officials that suffer the consequences if the decision turns out to be detrimental or unpopular." But when the Federal Government compels the States to take unpopular actions, "it may be state officials who will bear the brunt of public disapproval, while the federal officials who devised the regulatory program may remain insulated from the electoral ramifications of their decision." For this reason, federal officeholders may view this "departur[e] from the federal structure to be in their personal interests . . . as a means of shifting responsibility for the eventual decision." And even state officials may favor such a "departure from the constitutional plan," since uncertainty concerning responsibility may also permit them to escape accountability. If a program is popular, state officials may claim credit; if it is unpopular, they may protest that they were merely responding to a federal directive.[104]

This passage is hard to read with a straight face. The Court was, after all, deliberating over the fate of a law universally known as "Obamacare."[105] That inconvenient datum might be taken to cast doubt on the suggestion that the Constitution must be interpreted to proscribe federal spending programs that exert too much pressure on the states lest federal officials escape

---

101. 505 U.S. 144, 168 (1992).

102. Nat'l Fed'n of Indep. Bus. v. Sebelius, 132 S. Ct. 2566, 2660 (2012) (joint opinion).

103. *See supra* note 47 and accompanying text.

104. *Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. at 2660–61 (joint opinion) (internal citations omitted).

105. The reasoning is also at least somewhat hard to take from avowed originalists. For those scoring at home, the "blurred accountability" principle represents structural reasoning. Insofar as the joint opinion's embrace of the anti-compulsion principle rests on this rationale, it is not obviously justified by ordinary meaning originalism and therefore requires more elaboration than most originalists have provided regarding the relationship between structural principles or implications and a text's public meaning. For my critiques of originalism, see generally Mitchell N. Berman, *Originalism Is Bunk*, 84 N.Y.U. L. REV. 1 (2009), and Mitchell N. Berman, *Reflective Equilibrium and Constitutional Method: Lessons from John McCain and the Natural-Born Citizenship Clause*, *in* THE CHALLENGE OF ORIGINALISM: THEORIES OF CONSTITUTIONAL INTERPRETATION 246 (Grant Huscroft & Bradley W. Miller eds., 2011).

accountability for an unpopular law. But even if we abstract from the context of utterance, the claim should still strike us as resting upon a model of political accountability that is almost breathtakingly naïve.

It's not that I think there is nothing to this bit of political science wisdom. There just isn't enough to justify a flat rule that conditions on federal-spending grants to the states exceed Congress's power if they leave states "no real choice" other than to accept. The problem arises from the fact that the majority's anti-compulsion rule marks off for special, disfavored treatment the polar case while permitting adjacent cases on the relevant continuum. On the majority's approach, Congress is fully entitled to attach conditions to its spending programs that exert so much pressure on the states as to make it, let us say, very hard for state officials to decline. But as soon as the magnitude of pressure that a conditional offer exerts crosses the magical line that separates "pressure" from "compulsion," "voluntary" from "involuntary," and "really hard choice" from "no choice," the offer is invalid.

This is an implausible place to draw a constitutional line. To be sure, courts must routinely craft doctrine that attaches dichotomous consequences to phenomena that lie on either side of a largely arbitrary dividing line. The problem here is not, then, that the majority's line is arbitrary. The problem is that the line is perverse.

On any remotely realistic picture of American political and electoral dynamics, a federal offer that gives states "no choice" but to accept threatens accountability *less* than does an offer that puts substantial pressure on the states while leaving them some choice in the matter.[106] In the former case, it is much easier for a modestly informed voter to realize that the policy she dislikes was forced upon the states and therefore is the responsibility of federal agents.[107] In the latter, it will require vastly more sophistication for the voter to develop an informed view regarding whether the pressure was such that, all things considered, the state agents should or should not have acquiesced.

The authors of the joint opinion deem it "unmistakably clear. . . . that every State would have no real choice but to go along with the Medicaid Expansion."[108] They're right: it *is* unmistakably clear.[109] That's why a perfectly sensible concern with blurred lines of political accountability cannot justify the rule they defend. That concern would more sensibly *permit* congressional action at the extremes—either straightforward

106. For a brief summary of other criticisms of the Supreme Court's accountability theory, including citations to other authors, see Neil S. Siegel, *Commandeering and Its Alternatives: A Federalism Perspective*, 59 VAND. L. REV. 1629, 1632–33 (2006).

107. *Cf. id.* at 1632 (suggesting that the accountability issues of commandeering are exaggerated because engaged citizens are able to track the level of government responsible for particular initiatives).

108. *Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. at 2662 (joint opinion).

109. But perhaps I should say that it *seems* unmistakably clear yet might not be. *See supra* note 54.

commands to the states or inducements so weak as to be accepted only by the wholehearted—and *prohibit* or constrain offers that alter the option sets faced by state offerees in ways too complicated and subtle for voters to intelligently assess. Of course, I am not advocating that conditional spending offers that exert significant pressure short of compulsion should be prohibited. My point is only that the consideration on which the majority Justices would rely does not carry them where they wish to go.

III. Roberts, Once More

If, as Part II argues, it makes little sense to interpret our Constitution to prohibit Congress from using offers of federal funds to compel states to accede to conditions that Congress could not mandate, that does not cast doubt on the "anti-coercion principle." That principle could be construed as one that prohibits Congress from using offers of federal funds to coerce states to accede to conditions that Congress could not mandate. Little argument is necessary to establish that *this* is a sound principle. It follows from what I have elsewhere termed "the threat principle": ordinarily, if it is wrong to φ, it is wrong to threaten to φ.[110] And few people are ever moved to contest that principle.[111]

No, the objection to interpreting the Spending Clause as circumscribed by a true anti-coercion principle is not that Congress should be free to coerce the states into accepting behavioral conditions that Congress could not mandate. The objection is that that constraint is essentially meaningless. Because the states are not constitutionally entitled to federal funds,[112] the threat to withhold them is never a threat to act wrongfully, *constitutionally speaking*, therefore threats to withhold federal funding from states can never constitute any type of coercion that is constitutionally cognizable. Consequently, the objection continues, a constitutional limitation on

---

110. *See* Mitchell N. Berman, *Blackmail*, *in* THE OXFORD HANDBOOK OF THE PHILOSOPHY OF CRIMINAL LAW 37, 39 (John Deigh & David Dolinko eds., 2011). The "paradox of blackmail" is the criminal law counterpart to the puzzle of unconstitutional conditions: both ask how it can be wrongful within a particular normative system to threaten what would be permissible, within that system, to do (i.e., withhold a governmental benefit, disclose an embarrassing secret). My solutions to the two puzzles are analogous. In my view, Bill Edmundson is mistaken to assert that blackmail demonstrates that a proposal can be wrongfully coercive for reasons that do "not derive from or depend upon the wrongness of the declared unilateral plan [i.e., the conduct threatened]." Edmundson, *supra* note 37, at 457. In both cases, the *seemingly* permissible conduct threatened may be *im*permissible when undertaken for certain reasons, and the fact of the conditional offer might have evidentiary bearing on whether those reasons are present.

111. Nuclear deterrence is not a counterexample. If, as most people believe, it is morally permissible, all things considered, to threaten nuclear retaliation against a nation that launches an offensive nuclear attack, that is not because the threat does not constitute the moral wrong of coercion. It is because exceptional circumstances might justify engaging in the wrong of coercion, as is true of most moral wrongs, and because deterring nuclear attack provides adequate justification.

112. *See Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. at 2630 ("States have no entitlement to receive any Medicaid funds; they enjoy only the opportunity to accept funds on Congress' terms.").

Congress's spending power that is fairly described in "anti-coercion" terms must proscribe more than coercion.[113] I will argue in the next Part that this objection rests on a mistaken premise. It *can* be constitutionally wrong to withhold funds to which a state is not constitutionally entitled, and therefore can be constitutionally wrong—the wrong of coercion—to conditionally threaten to withhold them.

But this is getting ahead of ourselves. Here I aim only to bolster doubts raised in the previous Part about the coherence or soundness of the anti-compulsion principle as applied to federal spending programs by showing that the Chief Justice does not endorse that principle as unambiguously as a first read of his opinion suggests. I have already said that the Roberts opinion appears to maintain, with the joint opinion, that the Medicaid expansion is unconstitutional because it gives states "no choice" but to accept.[114] That is, the condition is impermissible, in Roberts's estimation, precisely because it amounts to impermissible compulsion. Yet several passages in Roberts's opinion indicate ambivalence on his part regarding whether the fact that a conditional spending offer by the federal government would compel state acceptance is sufficient to render the proposal unconstitutional.

A. The Modification Mystery

The first puzzle arises from Roberts's evident concern with whether the Medicaid expansion is a modification of the preexisting Medicaid program or, instead, a new program. Rejecting Justice Ginsburg's suggestion that "existing Medicaid and the expansion dictated by the Affordable Care Act are all one program simply because 'Congress styled' them as such,"[115] the Chief Justice's opinion concludes that the Medicaid expansion is in fact a new program, largely on the grounds that it "accomplishes a shift in kind, not merely degree."[116] The dissent strenuously disagrees.[117] Put aside for the moment who's right. The mystery is simply that Roberts should care. If, as Roberts appears to maintain, the dispositive constitutional question is whether the states had a "real choice" regarding whether to accept the

113. *Cf.* Kathleen M. Sullivan, *Unconstitutional Conditions*, 102 HARV. L. REV. 1413, 1456 (1989) ("There is good reason to turn elsewhere in a search for the rationale of unconstitutional conditions doctrine, both because the necessary baselines are elusive, once government benefits in this context are conceded to be gratuitous, and because government, which differs significantly from any given individual, can burden rights to autonomy through means other than coercion. Coercion thus begins rather than ends the inquiry.").

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

115. *Nat'l Fed'n of Indep. Bus.*, 132 S. Ct. at 2605 (quoting *id.* at 2635 (Ginsburg, J., dissenting in part)).

116. *Id.* at 2605 (Roberts, C.J.).

117. *Id.* at 2635–36, 2639–41 (Ginsburg, J., dissenting in part).

Medicaid expansion, it is not at all clear why the conclusion would differ depending on how that expansion is "properly viewed."[118]

Seemingly, the answer is this: When enacting the original Medicaid provisions, Congress had expressly reserved "[t]he right to alter, amend, or repeal any provision" of the statute.[119] Therefore, if the Medicaid expansion effected by the ACA was properly deemed an "amendment" to the preexisting Medicaid program, then notice of its possibility is fairly attributed to the states. But if the expansion were not an amendment, then the Court could conclude that it wasn't foreseeable. Attributing just this rationale to the Chief Justice, Justice Ginsburg asserts—without contradiction—that his claim that "the expansion was unforeseeable by the States when they first signed on to Medicaid" constitutes one of "three premises, each of them essential to his theory."[120]

But this is no solution to the mystery at all. Congress cannot, simply by reserving the right to amend a program, manufacture the authority to create amendments that exceed its constitutional authority. As the states rightly objected, the federal government's heavy reliance on its reservation of rights "confuses foreseeability and coercion."[121] The relevant question "is not whether States had any warning that Congress might exploit their dependence on Medicaid funding to coerce compliance with a massive expansion of the program, but whether Congress's coercive action is permissible."[122] If it is not permissible because it compels acceptance by giving states no choice other than to acquiesce, then the fact that the states can be held to have seen it coming is of no moment, for what they should also have seen coming is a judicial invalidation of the effort.

We can view the same point through a slightly different lens—through Roberts's intimation that it would have been permissible for Congress to accomplish exactly what it attempted through the Medicaid expansion had it first repealed the preexisting Medicaid program in its entirety and then enacted a new law that consisted of the prior law plus the Medicaid expansion. Justice Ginsburg takes the permissibility of this gambit for granted, framing the question that the Medicaid expansion presents around just that assumption: "To cover a notably larger population, must Congress take the repeal/reenact route, or may it achieve the same result by amending existing law?"[123] Again, Roberts does not deny this is so. To the contrary, his brief footnote response—that, due to practical or political considerations,

118. *Id.* at 2605 (Roberts, C.J.).

119. 42 U.S.C. § 1304.

120. *Nat'l Fed'n of Indep. Bus.*, 132 S.Ct. at 2630 (Ginsburg, J., dissenting in part).

121. Brief of State Petitioners on Medicaid at 41, *Nat'l Fed'n of Indep. Bus.*, 132 S.Ct. 2566 (No. 11-400).

122. *Id.* at 42.

123. *Nat'l Fed'n of Indep. Bus.*, 132 S.Ct. at 2629 (Ginsburg, J., dissenting in part).

repeal and reenactment "would certainly not be that easy"[124]—strongly implies his agreement that it would be constitutional in the unlikely event it were to occur. Again, though, it is mysterious why this should be if the constitutionally relevant inquiry is whether states have a realistic option to reject Congress's proposed deal.

B. The Reasons Riddle

For a second puzzle, consider Roberts's curious response to the states' "claim that this threat serves no purpose other than to force unwilling States to sign up for the dramatic expansion of health care coverage effected by the Act."[125] "Given the nature of the threat and the programs at issue here," he observed,

> we must agree. We have upheld Congress's authority to condition the receipt of funds on the States' complying with restrictions on the use of those funds, because that is the means by which Congress ensures that the funds are spent according to its view of the "general Welfare." Conditions that do not here govern the use of the funds, however, cannot be justified on that basis. When, for example, such conditions take the form of threats to terminate other significant independent grants, the conditions are properly viewed as a means of pressuring the States to accept policy changes.[126]

The response appears to maintain that Congress's purposes or reasons for action are constitutionally relevant and that, in enacting the Medicaid expansion, Congress was motivated by bad ones.[127] Yet if, as the anti-compulsion rendering of the anti-coercion principle appears to have it, a conditional offer exceeds Congress's power just because it leaves the states with "no choice" but to accept, it is something of a riddle why Congress's purposes should matter. If compulsion is the constitutional wrong, and if the ACA's threat to withhold all Medicaid funding unless the recipient state agrees to cover a new class of beneficiaries does in fact "force unwilling States to [accede to that condition]," it should be neither here nor there that the threat "serves no purpose other than" to secure compliance.

And why does it matter whether "the conditions are properly viewed as a means of pressuring the States to accept policy changes"? We know from *Steward Machine*, by way of *Dole*, that pressure by itself does not constitute compulsion.[128] So one might have thought, consistent with the body of Roberts's opinion, that how the conditions "are properly viewed" is again

---

124. *Id.* at 2606 n.14 (Roberts, C.J.).

125. *Id.* at 2603.

126. *Id.* at 2603–04.

127. *See id.* at 2606–07 (Roberts, C.J.) ("Congress may not simply conscript state agencies into the national bureaucratic army, . . . and that is what it is attempting to do with the Medicaid expansion.") (citations and internal quotation marks omitted).

128. *See supra* note 13 and accompanying text.

irrelevant. If the conditions are properly viewed "as a means of pressuring the States," but the pressure exerted leaves the states with some choice in the matter, then the condition does not produce compulsion and is constitutional. Contrariwise, if the pressure exerted leaves the states with "no choice" in the matter, then we would have compulsion, hence unconstitutionality, even if the conditions are "properly viewed" in some other light.

In short, this passage appears to consider it relevant to the constitutional inquiry—perhaps, indeed, fully inculpatory—that Congress's "purpose" behind this particular threat to withhold a benefit was to "pressure" reluctant states into behaving in a manner that Congress could not mandate. But it is not clear why, on an unadorned anti-compulsion construal of the governing constitutional principle, Congress's reasons for acting should be relevant.

C. *The Penalty Puzzle*

A final puzzle attaches to Roberts's tantalizing but underdeveloped suggestion that withholding a benefit to which a state is not constitutionally entitled is unconstitutional if non-provision of the benefit would penalize the exercise of a state's constitutional prerogatives. "Nothing in our opinion," concludes the Chief Justice near the end of his spending power analysis,

> precludes Congress from offering funds under the Affordable Care Act to expand the availability of health care, and requiring that States accepting such funds comply with the conditions on their use. *What Congress is not free to do is to penalize States that choose not to participate in that new program by taking away their existing Medicaid funding.*[129]

This short passage provokes at least two questions. First, what does it mean to "penalize" a state (or to impermissibly "penalize" a state)? Presumably "to penalize" is equivalent to "to impose a penalty." So we could rephrase the question: What is a "penalty"? Second, what is the relationship between penalties and compulsion?

We should not have to travel far for an answer to the first query. As luck would have it, resolution of the taxing power question turned precisely on the Court's answer to the question of whether the provision that required citizens who fail to secure minimum health insurance coverage to pay a sum to the IRS levied a "tax" or imposed a "penalty."[130] Justices Scalia, Kennedy, Thomas, and Alito, in dissent on this point, concluded the latter.[131] Chief Justice Roberts, in a portion of his opinion joined by the remaining justices, concluded the former.[132] Whether one concludes that the putative tax was or was not a "penalty," surely one must first know what it is for an

129. *Nat'l Fed'n of Indep. Bus.*, 132 S.Ct. at 2607 (Roberts, C.J.) (emphasis added).
130. *Id.* at 2594–600.
131. *Id.* at 2652–55 (joint opinion).
132. *Id.* at 2600 (Roberts, C.J.).

exaction to be a penalty. And Roberts is quick to endorse the definition offered by past cases: "[I]f the concept of penalty means anything, it means punishment for an unlawful act or omission."[133] In full accord on the definitional point, the joint opinion declares that "[o]ur cases establish a clear line between a tax and a penalty: A tax is an enforced contribution to provide for the support of government; a penalty . . . is an exaction imposed by statute as punishment for an unlawful act."[134]

Unfortunately, whatever the merit to this conceptualization of penalty in the tax context, it cannot be what Roberts has in mind when charging that the Medicaid expansion impermissibly threatens to penalize non-acquiescing states. The state challengers to the provision argue, and a Supreme Court majority agrees, that conditions on the new Medicaid funds transgress the "anti-coercion principle" because non-participation in the new program is not a realistic option.[135] But nobody argues, and it is not plausible, that the Medicaid expansion makes non-participation in the new program "unlawful." So if, through the Medicaid expansion, Congress is threatening "to penalize states that choose not to participate in that new program," it must be the case that the withdrawal of benefits to which a state is not constitutionally entitled can constitute a penalty even when the withdrawal is predicated on something other than "an unlawful act or omission" by the state. The problem is that Roberts offers the reader no clue, outside this brief and enigmatic passage, regarding what concept of penalty he means to employ.

Actually, the problem runs deeper if we are to insist that whatever conception of penalty Roberts might have dimly in mind must fit with the anti-compulsion reading of his opinion. To see why such fit is doubtful, imagine (contrary to fact, I am willing to assume) that a state's existing Medicaid funding, though substantial, were not so great that the state could not reasonably choose to forgo it as the price for not accepting the Medicaid expansion. Imagine too that a state were to exercise its practical option to say no and that, as a result, Congress were to take away all its Medicaid funding. How should we analyze the case?

Three possible characterizations of Congress's action seem most eligible: (1) withdrawal of funding under these circumstances does not "penalize" the affected state, and is permissible; (2) withdrawal of funding under these circumstances does "penalize" the affected state, and is impermissible; and (3) withdrawal of funding under these circumstances does "penalize" the affected state, but is nonetheless permissible. (The fourth logical possibility—that withdrawal of funding does not "penalize" the state, and is impermissible—is a nonstarter.) None of these possibilities sits well with an unadorned "anti-compulsion" reading of the Chief Justice's opinion.

---

133. *Id.* at 2596 (internal quotation marks omitted) (quoting *United States v. Reorganized CF & I Fabricators of Utah, Inc.*, 518 U.S. 213, 224 (1996)).

134. *Id.* at 2651 (joint opinion) (internal quotation marks and citations omitted).

135. *Id.* at 2662–64 (joint opinion).

Possibility (1) is not attractive, for it makes the presence of compulsion constitutive of whether an exaction is a penalty, yet it is commonplace that there are some exactions properly denominated penalties that one could rationally choose to incur. Possibility (3) is also not attractive because it seems to make penalty analysis do no work at all. On reading (3), Roberts should not have said (as he did) that Congress is not free to penalize states that choose not to participate in the new program; he should have said only that Congress is not free to compel states into participating. That leaves possibility (2). How plausible this proposition is must await further analysis of the concept of penalty. But if it does prove plausible, it creates tension with the view that, for Roberts, it is decisive that rejecting the Medicaid expansion was not a real or realistic option for the states. On possibility (2), Roberts would be saying not only (or not even) that Congress may not compel states to accept the Medicaid expansion, but also (or rather) that Congress may not threaten to penalize states that don't.

* * *

To summarize, Chief Justice Roberts's opinion, written for himself and for Justices Breyer and Kagan, provokes at least three questions. First, what is the relevance of the fact, if true, that the Medicaid offer was not a modification of the preexisting Medicaid program? Second, what is the constitutional significance of Congress's reasons for structuring the Medicaid expansion as it did? Third, what is a "penalty" for Spending Clause purposes (or more generally), and how does the concept of penalty interact with those of compulsion and coercion?

The presence of these puzzles demonstrates that Roberts, Breyer, and Kagan might well have harbored doubts—doubts wholly consistent with the equivocal language used in Spending Clause precedents—about the "straightforward" anti-compulsion reading of the "anti-coercion principle" favored by the joint opinion. But they do more than that. As we will see, these puzzling aspects of the opinion lend support to (I do not say they "compel") the interpretations of coercion and penalty that I offer in the next Part.[136]

---

136. In a careful and thorough analysis of the *NFIB* Spending Clause holding, Sam Bagenstos maintains that my analysis of conditional spending confronts "two significant problems." Samuel R. Bagenstos, *The Anti-Leveraging Principle and the Spending Clause After* NFIB, 101 GEO. L.J. (forthcoming 2013) (manuscript of August 2012 at 34). The second is that it has overly broad implications. I address this worry as the fifth objection in Part V, *infra*. The first problem is that my analysis "would not be an attractive interpretation of the Roberts opinion," in part because whereas I end up concluding that *Dole* was wrongly decided, the Chief Justice accepts it unquestioningly. *Id.* at 38.

Given this criticism, I should make very clear that I do not claim that my analysis is an interpretation of the Roberts opinion. While I do claim that features of his opinion cohere well with my analysis, I fully agree with Bagenstos that my analysis is inconsistent with some things that the Chief Justice says. Endeavoring to make better sense of that opinion than my analysis does, Bagenstos advocates (somewhat half-heartedly) what he calls "the anti-leveraging principle," which provides that "[w]hen Congress takes an entrenched federal program that provides very large sums

IV. The Medicaid Expansion and the Anti-Coercion Principle, Rightly Understood

Readers sympathetic to the Medicaid expansion are likely to find the arguments to this point, if persuasive, heartening. If what I have argued so far is correct, the Medicaid provisions of the ACA should be held unconstitutional only if the consequence that the biconditional proposal threatens to impose on a non-accepting state—withdrawal of all Medicaid funding—would be unconstitutional. And common wisdom holds that this cannot be. As the amicus brief for former Surgeon General David Satcher and others maintains, "[f]or the financial inducement offered by Congress to become unconstitutionally coercive, that inducement must, at a minimum, deprive the state of something *to which the state is otherwise entitled*."[137] And, the argument continues, no state is entitled to federal Medicaid funds.[138] I argue in this Part that the major premise is mistaken.

My argument proceeds in three steps. Subpart IV(A) formulates and provisionally defends a general principle concerning one likely entailment or corollary of constitutional rights. I call this the "anti-penalty principle" mostly because it is apt, but also because that designation makes for a pleasing companion to the anti-coercion principle we have already been discussing. The next two steps assess the Medicaid expansion in light of this general principle. Subpart IV(B) introduces a highly stylized or schematic understanding of the Medicaid expansion as consisting of three discrete conditional offers: an offer of funds for the blind, the disabled, the elderly, and poor families with dependent children; an offer of funds for adult childless poor; and an offer to render states eligible for the first offer only if they accept the second. It concludes that, if this is how the Medicaid expansion is fairly or properly viewed, it runs afoul of the anti-penalty principle and, as a consequence, of the anti-coercion principle too. Subpart IV(C) considers whether the conclusion from subpart IV(B) changes when we recharacterize the Medicaid expansion as a single program or package.

---

to the states and tells states that they can continue to participate in that program only if they *also* agree to participate in a separate and independent program, the condition is unconstitutionally coercive." *Id.* at 5. The warrant for this principle is that only it makes sense of all aspects of Chief Justice Roberts's opinion. A central difference between my approach and Bagenstos's, accordingly, concerns just how seriously each of us takes all aspects of that opinion. As I read Bagenstos, he appears to assume that Roberts's opinion is fully coherent and well thought out. I, in contrast, read the opinion as gestural and at least partly inchoate. Divining normatively defensible parameters for the exercise of Congress's spending power is a real challenge. In light of both the difficulty of the problem and the somewhat meandering tenor of Roberts's opinion, I find it more plausible and profitable than does Bagenstos to understand that opinion to be grasping toward a solution rather than to have captured one.

137. Brief for David Satcher et al. as Amici Curiae Supporting Respondents at 2, Florida v. U.S. Dep't of Health and Human Servs., 132 S.Ct. 2566 (2012) (No. 11-400), 2012 WL 588459, at *2 (emphasis added).

138. *Id.*

Although I acknowledge that this is a difficult question, I conclude that, on the facts of this case, it probably does not.

A. The Anti-Penalty Principle

1. Introduction to Unconstitutional Conditions.—We saw in Part III that there exists what the Justices described as a well-established definition of "penalty": a penalty is an exaction imposed as punishment for unlawful conduct.[139] That well-settled definition, however, is localized. It is the definition accepted in the tax context for distinguishing exactions that are and are not permissible exercises of Congress's taxing power: the exaction is permissible if a "tax," impermissible if a "penalty."[140] It does not apply across the legal board. In particular, courts have frequently used or gestured to a very different conception of penalty in "unconstitutional conditions" cases.[141]

Although courts and commentators often refer to the "unconstitutional conditions doctrine," if a doctrine is a set of rules or tests, then there is no such doctrine—at least none with more than trivial content.[142] Better to think and speak of a "conditional offer problem" or a "conditional offer puzzle"—the difficulty of properly analyzing governmental offers of benefits that it is not constitutionally obligated to provide conditioned on the offeree's waiver or non-exercise of a constitutional right. Federal offers of funds to states on the condition that they exercise their sovereign prerogatives in any fashion that Congress could not mandate raise the conditional offer problem. So too do countless offers of benefits conditioned on the waiver of individual rights: welfare grants for the poor conditioned on their agreement to be subjected to warrantless, suspicionless searches,[143] subsidies for public broadcasters conditioned on their agreement not to editorialize,[144] lower criminal sentences conditioned on a defendant's waiver of his right to put the state to

---

139. *See supra* subpart III(C).

140. Nat'l Fed'n of Indep. Bus. v. Sebelius, 132 S. Ct. 2566, 2651 (2012) (joint opinion).

141. *See, e.g.*, Wyman v. James, 400 U.S. 309, 340 (1971) (stating that the right to welfare benefits conditioned on warrantless searches amounts to a civil penalty).

142. You could read a dozen scholarly discussions of "the unconstitutional conditions doctrine" before running into a clear statement of what the doctrine is supposed to say or what its content is. When a statement of the doctrine's content is provided, it often goes something like this: "Essentially, this doctrine declares that whatever an express constitutional provision forbids government to do directly it equally forbids government to do indirectly." William W. Van Alstyne, *The Demise of the Right-Privilege Distinction in Constitutional Law*, 81 HARV. L. REV. 1439, 1445–46 (1968). Courts have, on occasion, said such things. But I'd be surprised if anybody in a generation has believed that broad claim to be true, which suggests that it could be an accurate rendition of the doctrine only if everybody believed that the "unconstitutional conditions doctrine" is false. Not everybody does, so it must have different content.

143. *E.g.*, *Wyman*, 400 U.S. 309 (1971); Sanchez v. County of San Diego, 464 F.3d 916 (9th Cir. 2006).

144. FCC v. League of Women Voters, 468 U.S. 364 (1984).

its burden of proof,[145] land use variances conditioned on a landowner's grant to the public of some of its property rights,[146] and on and on. Since the earliest cases that first self-consciously identified the conditional offer problem, way back in the 1870s,[147] courts have failed so spectacularly to analyze the problem in a coherent or even consistent fashion as "to make a legal realist of almost any reader," as Seth Kreimer aptly put it.[148] The only rendering of the "unconstitutional conditions doctrine" that is remotely faithful to the cases would maintain that sometimes conditional offers of the foregoing sort are permissible, while sometimes they aren't.[149]

Be that as it may, courts have, predictably, experimented with a variety of analytic approaches. And one of the more common turns on the concept of a penalty. The idea, very simply (perhaps a little too simply, as we will see), is that it is unconstitutional to penalize the exercise of a constitutional right.[150] Call this succinct claim "the anti-penalty principle" (AP). It is defeasible. Put in familiar terms, penalizing a constitutional right infringes but does not violate the right. Thus:

AP: It is presumptively unconstitutional for the government to penalize the exercise of a constitutional right.

Furthermore, by dint of the straightforward idea that it is impermissible to threaten what it is impermissible to do (the heart of a true anti-coercion principle), it is also presumptively unconstitutional to *threaten* to penalize the exercise of a constitutional right.

Judicial statements that endorse AP or something very close to it are common. We have already seen, for example, that Justices Scalia and Thomas approved it in *Lee v. Weisman*[151] and that the Chief Justice at least flirts with it in *NFIB*.[152] We also observed that, for this proposition to be useful, we will need to know what "penalty" and "penalize" mean—something that the frequent judicial endorsements of AP rarely divulge. One might reasonably complain, therefore, that AP is not, by itself, terribly informative. But even if not as informative or fully specified as we'd like, I anticipate that most readers, likely operating with just an inchoate sense of what a penalty is, will find it rather intuitive. Quickly: May the state

145. *E.g.*, Bordenkircher v. Hayes, 434 U.S. 357 (1978).

146. *E.g.*, Dolan v. City of Tigard, 512 U.S. 374 (1994); Nollan v. Cal. Coastal Comm'n, 483 U.S. 825 (1987).

147. The earliest cases involved state laws that conditioned the grant of corporate privileges on an out-of-state corporation's agreement not to remove suits filed against it to federal court. For a discussion, see Berman, *supra* note 19, at 59–70.

148. Seth Kreimer, *Allocational Sanctions: The Problem of Negative Rights in a Positive State*, 132 U. PA. L. REV. 1293, 1304 (1984).

149. *Id.*

150. For discussion and analysis of this principle in the case law, see Sullivan, *supra* note 113, at 1433–43.

151. *See supra* subpart II(B).

152. *See supra* subpart III(C).

withdraw eligibility for free school lunches from the children of mothers who obtain abortions? Surely not. And why not? Because doing so impermissibly penalizes a woman's exercise of her constitutional right to an abortion.[153]

I think we are therefore entitled to embrace AP as a working hypothesis—a hypothesis, I emphasize, not a conclusion.[154] The goal for this section, accordingly, is to develop conceptions of "penalty" and "penalize" pursuant to which AP is in fact true. Moreover, because AP is so frequently invoked in an effort to explain why some conditional offers of "benefits" (i.e., largesse, advantages, or things of value to which the beneficiary is not constitutionally entitled) are unconstitutionally coercive, we hope further for a definition of penalty that will capture at least some withdrawals or denials of benefits. In short, the desiderata for a definition of penalty are (1) that it render AP true and (2) that it encompass at least some failures or refusals to furnish benefits (as just defined).

2. The Baseline Problem.—Here's a first stab, courtesy of the Court's Self-Incrimination Clause jurisprudence tracing back to its 1965 decision in *Griffin v. California*.[155] The Fifth Amendment privilege against self-incrimination, the Court has consistently held, permits defendants not only to remain silent, free from criminal punishment, but also "to suffer no penalty

153. An alternative explanation would be that the state is impermissibly trying to discourage women from exercising their right to an abortion. But this is a bad explanation. It is true that the state is prohibited from trying to influence exercises of some rights. For example, it may not act for the purpose of encouraging or discouraging attendance at houses of religious worship. But this is not true of all rights, and, as a matter of positive law, the state is permitted to try to encourage women to "choose life." You might think that is a mistaken decision. Maybe it is, maybe it isn't. The critical point is that the hypothetical action described in the text should strike us as plainly unconstitutional even assuming arguendo that the state is constitutionally permitted to try to discourage women from exercising their right to an abortion. That is, the state may not try to discourage women from having abortions by the particular means of threatening to penalize them if they do.

154. At least one eminent commentator has denied this. Cass Sunstein once went so far as to conclude that "[t]he Constitution offers no general protection against the imposition of penalties on the exercise of rights." Cass R. Sunstein, *Why the Unconstitutional Conditions Doctrine Is an Anachronism (With Particular Reference to Religion, Speech, and Abortion)*, 70 B.U. L. REV. 593, 603 (1990) [hereinafter Sunstein, *Anachronism*]. That is, he flatly denied AP. He could maintain this position, however, only because he already accepted a definition of penalty (a non-normative one, to jump ahead) according to which AP is false. The other possibility is to accept AP and then try to formulate a conception of penalty that vindicates it. I think the second approach far preferable because most of us start with a strong (though necessarily defeasible) pretheoretical commitment to AP.

Revealingly, when he later converted his 1990 article into a book chapter, Sunstein softened his rejection of an anti-penalty principle. The claim then became that "*[i]t is not clear* that there is any general protection, in the Constitution, against penalties on rights." CASS R. SUNSTEIN, THE PARTIAL CONSTITUTION 300 (1993) (emphasis added). That is a very different claim, for indeed it is *not* "clear" that AP is true. It is only "likely" or "intuitively plausible." The task is to see whether "penalty" can be specified in a manner that vindicates AP. As it turns out, the specification that I will propose is different from that which Sunstein assumes. *See infra* note 177.

155. 380 U.S. 609 (1965).

. . . for such silence."[156] And "penalty," the Court has emphasized, "is not restricted to fine or imprisonment. It means . . . the imposition of any sanction which makes assertion of the Fifth Amendment privilege 'costly.'"[157] Given its embrace of the anti-penalty principle, and consistent with its understanding of "penalty" as governmental conduct that makes exercise of the right more costly, the Court has prohibited, for example, the prosecution from commenting on the accused's silence,[158] the court from instructing jurors that silence is evidence of guilt,[159] and the organized bar from sanctioning non-testifying attorneys.[160]

Griffin provides a starting point, but its use of the adjective "costly" cannot stand without qualification. The Constitution does not plausibly forbid actions that make exercise of Fifth Amendment rights costly in some abstract or objective sense. The underlying notion must be comparative. Let us then read *Griffin*'s definition of penalty to cover actions that make individual conduct "*more* costly." For this definition to be useful, we need as well an answer to the question "more costly than *what*?" The "what" is standardly termed the "baseline." Thus do we have the following proposed definition of penalty and penalize:

P: Any governmental act or omission, G, penalizes (i.e., imposes a penalty upon) some conduct, C, by an actor, A, if G makes C more costly for A than C would have been for A [had the appropriate baseline state of affairs obtained].

So far, so good. But not far enough. Plainly, we need to replace the bracketed language with a specification of the appropriate baseline.

Although, in principle, any number of conceivable baselines might be identified, most or all will fall into one of two classes: either normative or non-normative.[161] A normative baseline is constituted by the treatment that the agent, A, *should* get.[162] Non-normative baselines fall into at least two subclasses: positive and counterfactual.[163] A positive baseline is constituted by some actual state of affairs, such as the state of affairs that A in fact enjoyed prior to the governmental act or omission in question (a historical

156. Malloy v. Hogan, 378 U.S. 1, 8 (1964).

157. Spevack v. Klein, 385 U.S. 511, 515 (1967) (quoting *Griffin*, 380 U.S. at 614).

158. *Id.*

159. *Id.*

160. *Id.* at 514.

161. To simplify the discussion, I will put aside possible combinations of normative and non-normative baselines. This is a legitimate simplification because my goal in this section is to present sympathetically the objections that scholars have raised *against* efforts to solve the conditional offer problem by invoking the anti-penalty principle. Complicating the menu of possible baselines is a move for proponents, not opponents, of AP.

162. Kreimer, *supra* note 148, at 1373–74.

163. *See id.* at 1359, 1363, 1371 (identifying three types of baseline: equality, history, and prediction, the latter of which correspond to the non-normative positive and counterfactual baselines); *see also* Berman, *supra* note 19, at 13 (identifying Kreimer's equality, history, and prediction baselines as "positive ('history' and 'prediction') and normative ('equality')").

baseline) or the state of affairs that agents similarly situated to A enjoy in the jurisdiction or elsewhere (a comparative baseline).[164] A counterfactual baseline is constituted by what the world would look like under some specified counterfactual circumstance, such as the state of affairs that A would enjoy if the government were disabled from conditioning a benefit in the particular way that it has, and thus would have to provide it either more broadly or less broadly.[165]

With this thumbnail taxonomy of possible baselines in hand, we reach the difficulty that confronts proponents of a penalty-based solution to the conditional offer puzzle: "the baseline problem."[166] The supposed problem is that no non-normative baseline provides a specification of P pursuant to which AP is true, and no normative baseline provides a specification of P pursuant to which it encompasses any non-provision of "benefits." Thus, the baseline cannot be specified in any fashion that provides a definition of penalty that satisfies both of our stated desiderata: (1) that it renders AP true and (2) that it shows that at least some failures to provide benefits impermissibly penalize rights.

Let us take these two claims in order. Take the most obvious candidate for a non-normative baseline: the "historical" baseline. Fleshing out P by allowing history to constitute "the appropriate baseline" yields the following definition, that we can denominate P1:

P1: Any governmental act or omission, G, penalizes some conduct, C, by an actor, A, if G makes C more costly for A than C would have been for A prior to G.

P1 is a conceivable stipulative definition of "penalty." However, it is not a definition that makes AP true. An increase in postage rates makes many exercises of First Amendment rights more costly than they would be absent the increase. The decision to locate a polling place here rather than there makes it more costly for some people to exercise their right to vote. These and countless other governmental actions make exercise of rights more costly than they would be absent those actions, yet do not plausibly raise constitutional alarms. Again: We are not looking for just any definition of "penalty" that minimally comports with linguistic usage; we are hunting for a definition of "penalty" that makes AP true.

---

164. *See* Kreimer, *supra* note 148, at 1359–64 (examining "history as a baseline" and "equality as a baseline").

165. The seminal exploration of the types of baselines that might help solve the conditional offer problem is Kreimer, *supra* note 148. *See also, e.g.*, Kenneth W. Simons, *Offers, Threats, and Unconstitutional Conditions*, 26 SAN DIEGO L. REV. 289 (1989).

166. Sometimes the "baseline problem" is raised as a challenge for accounts of constitutional "penalties." More often, it arises in the context of assessing whether conditional offers of benefits can ever be wrongfully "coercive." Because the most common way in which a threat to withhold a "benefit" can constitute coercion will be that it penalizes the exercise of a constitutional right, these two formulations of the baseline problem are fundamentally the same.

If P1 does not fit the bill, here's a specification of P that does make AP true:

P2: Any governmental act or omission, G, penalizes some conduct, C, by an actor, A, if G makes C more costly for A than C would have been for A had A received that to which A is constitutionally entitled.

This is a somewhat complicated way to say that government penalizes conduct by treating the actor who engages in the conduct less well than the actor should be treated, constitutionally speaking. It is therefore the most natural reflection of a normative baseline. Unlike P1, P2 makes AP true—indeed, P2 makes AP tautological. But it makes AP true in a way such that AP cannot be violated by the withdrawal or nondisbursement of benefits, precisely because benefits are defined as goodies to which the beneficiary is not constitutionally entitled. P2 does not satisfy our second desideratum.

On the basis of reasoning like this, some of the leading constitutional theorists of our day have concluded that the withdrawal of benefits can never penalize rights in any sense of "penalizing rights" that is constitutionally suspect, which is also to say that *threats* to withdraw benefits (on failure of stated conditions) can never be unconstitutionally coercive. As Kathleen Sullivan concluded in an influential article, "To hold that conditions coerce recipients because they make them worse off with respect to a benefit than they *ought* to be runs against the ground rules of the negative Constitution on which the unconstitutional conditions problem rests."[167]

3. The Baseline Solution.—I believe that this scholarly near-consensus is mistaken. Its error is to suppose that the set of eligible normative baselines is exhausted by states of affairs describable without reference to government's reasons for causing them, or allowing them to obtain. The "penalty skeptics" (or "coercion skeptics") maintain that if an actor is not constitutionally entitled to be provided with a benefit, then it cannot be improper for the state to withhold it.[168] What they do not adequately appreciate is that government's reasons for actions might be constitutionally

167. Sullivan, *supra* note 113, at 1450; *see also, e.g.*, LOUIS MICHAEL SEIDMAN & MARK V. TUSHNET, REMNANTS OF BELIEF: CONTEMPORARY CONSTITUTIONAL ISSUES 79 (1996) ("Where the government has no obligation to provide the subsidy at all, it makes no one legally worse off by conditioning the subsidy on desired behavior. Under this test, however, the conditional-offer doctrine does no work."); Samuel R. Bagenstos, *Spending Clause Litigation in the Roberts Court*, 58 DUKE L.J. 345, 373 (2008). *See generally* Larry Alexander, *Understanding Constitutional Rights in a World of Optional Baselines*, 26 SAN DIEGO L. REV. 175 (1989).

168. A common formulation is that benefits can be withheld "for any reason or no reason at all." For a charming illustration that people often say such a thing without reflection, see Rankin v. McPherson, 483 U.S. 378, 383–84 (1987): "Even though McPherson was merely a probationary employee, and even if she could have been discharged for any reason or for no reason at all, she may nonetheless be entitled to reinstatement if she was discharged for exercising her constitutional right to freedom of expression." Obviously, McPherson's possible entitlement to reinstatement contradicts the supposition that she was dischargeable "for any reason."

relevant, such that the non-provision of a benefit to which a would-be beneficiary is not constitutionally entitled can be unconstitutional because of the reasons for which it is not provided.[169] Put in a familiar vocabulary, the skeptics focus exclusively on the outputs of state action, wholly neglecting the inputs.

If government's reasons for action (including inaction) are constitutionally relevant, then we should entertain the possibility that the non-provision of benefits is unconstitutional if motivated by bad reasons, and the task becomes one of identifying the reasons that count as bad. Here's a rough-cut proposal: government may not withhold benefits it would otherwise provide for the purpose either of discouraging agents from exercising their constitutional rights or of punishing them for doing so.[170] Stated differently, if government has reasons to provide a particular benefit to a particular potential beneficiary, it may not withhold that benefit in order to make the exercise of constitutional rights costly or painful. Let us try to convert these general thoughts into a definition of penalty, formulated as a specification of P:

> P*: Any governmental act or omission, G, penalizes (i.e., imposes a penalty upon) some conduct, C, by an actor, A, if G makes C more costly for A than C would have been for A had the government not undertaken G and if the government engaged in G, rather than not-G, for the purpose of making C more costly or painful.

Please take P* as a work in progress. Very likely, it can be improved upon. The core idea, again, is that if government could have made C less costly than it did make C, but did not choose that path because of—and not in spite of—its anticipation that its action would prove costly to A (presumably, for deterrent or punitive reasons), then its pursuit of the more costly-to-A path imposes a penalty on A's doing of C. Put in the language of reasons, *the state may not take the expected fact that a proposed course of action would make the exercise of rights more costly or more painful as a reason in favor of that course of action.* (More costly or more painful than *what*? More costly or more painful than would be the case if the state did otherwise.)

Two things about P* merit emphasis. First, it is not the case that, on this definition, all withholdings of benefits amount to a penalty. The University of Texas School of Law, a state actor, offers a faculty position to Lucy Taylor, conditioned on her agreement to teach tax. She declines, as is her constitutional right. In response, UT carries out its threat not to employ

169. This objection is not original to me. *See, e.g.*, John H. Garvey, *The Powers and Duties of Government*, 26 SAN DIEGO L. REV. 209, 224 (1989) (noting that it is often said that the government cannot withhold benefits for a bad reason).

170. By acting "in order to punish" or "for the purpose of punishing," I mean that one acts on vindictive or retributive non-instrumental reasons for imposing costs or hardship.

her. This non-provision of a benefit need not be tainted by any purpose that renders it a penalty. That withholding the job would make Taylor's exercise of her right not to teach tax more costly or painful need play no role in the Law School's deliberation. It may simply be that the Law School has inadequate affirmative reason to employ Taylor if doing so would not fill its curricular needs.[171]

Second, that some action by the state does penalize some conduct is not enough to render the state action suspect. Some conduct the state is entirely free to penalize. Criminal punishments are penalties on the proposed account, for the state imposes them to make the proscribed conduct more costly or painful than it would be otherwise, and does so for the purpose of discouraging and/or punishing it.[172] But they are unproblematic precisely to the extent that people do not have a right to engage in the conduct criminalized and thus penalized. The claim—represented by AP—is that *the state is obligated not to penalize the exercise of rights*. Part of what it is to have a constitutional right to φ is to have a right not to be penalized for φing—in the sense of penalty captured by P*.[173] Thus, combining the anti-penalty principle, AP, with P* as the specification of what a penalty is yields the following principle:

AP*: It is presumptively unconstitutional for the government to make the exercise of a constitutional right more costly for the right holder than it would be had the government acted otherwise where the government would have acted otherwise but for a purpose of making the exercise of the right more costly or painful.

I have now formulated the suggestion in a variety of ways that approximate one another even if they don't correspond precisely.[174] I have not yet provided argument for it. I do not believe that any slam-dunk argument in favor of it exists. For example, AP* cannot be deduced from incontrovertible first principles or even from principles that, if controvertible, are not in fact contested. The best argument for AP* must be largely coherentist: First, AP* is highly plausible on its face. Second, AP* best

171. I am assuming that the Law School would not want to hire Taylor if she could not, or would not, teach tax. But the case could be otherwise. The Law School might deem Taylor a sufficiently attractive candidate to warrant hiring her regardless of her curricular commitments. In that case, the conditional offer would be extended as a way to pressure or induce her to teach a subject that would make her yet more attractive. For a discussion of this possibility, see *infra* Part V, Objection 6. I am grateful to David Strauss for pressing me on just this point.

172. Again, see *supra* note 170 for what I mean by the "purpose" of "punishing."

173. Not all rights are rights to φ. I mean the claim in text to accommodate these other types of rights too.

174. One respect in which the formulations differ concerns how the bad purposes or reasons function in the state's deliberation—in particular, whether the reasons I identify turn the withholding of a benefit into a penalty only if they serve a but-for role, or if they serve any motivating function, or are "substantial factors," and so on. *See* Berman, *supra* note 19, at 27 & n.103. I leave this an open question for now.

accounts for widespread intuitions about a wide range of cases, actual and hypothetical, and for judgments about cases that, if not immediately intuitive, withstand critical scrutiny.[175]

The latter claim cannot be fully demonstrated in this Article, given the number and diversity of cases that a coherentist analysis would have to address. I have, however, taken a stab at the project elsewhere.[176] Here, I can proceed only some distance toward establishing the plausibility and attractiveness of AP*. As a first step, let us consider two hypothetical cases, what I will call *Vindictive Sentencing* and *Short Zoning*.

Vindictive Sentencing

Harris is convicted of robbery, a second-degree felony punishable under state law by a sentence of imprisonment from two to twenty years. Judge Davis imposes a sentence of ten years. Harris appeals his conviction on the grounds that a motion to exclude certain eyewitness testimony was improperly denied. The court of appeals agrees and vacates the conviction. Harris is convicted on retrial and once again comes before Judge Davis for sentencing. This time, however, Judge Davis imposes a sentence of twenty years. She explains this longer sentence in open court: "We simply cannot have guilty people challenging this court's orders with impunity."

Short Zoning

The three-member local land use commission is considering what zoning restrictions to impose on beachfront property. Commissioners Smith and Jones observe that height limits of forty feet would adequately serve the community's environmental and aesthetic interests. Commissioner Brown, speaking last, agrees. But he also observes that a thirty-foot limit would allow the Commission to extract concessions from homeowners in exchange for permission to build up to forty feet. "Good point," says Commissioner Smith. "Brilliant," adds Commissioner Jones. They vote unanimously to limit beachfront property to thirty vertical feet.

After the new zoning rules go into effect, the Johnsons, owners of a beachfront lot, seek a variance from the height restrictions that would allow

---

175. Note that I do not maintain that the conjunction of AP and P* *perfectly* accounts for widespread intuitions about a wide range of cases. Some of my conclusions with regard to actual cases differ from what the courts have held; some may differ from your own intuitions. The method of reflective equilibrium requires that we be willing to abandon some of those case-specific intuitions in order to produce a set of mutually supportive beliefs that we can accept on reflection better than any alternative set. For elaboration, see Berman, *supra* note 105, at 259–61. To be sure, if application of AP* yields conclusions that you are firmly convinced are mistaken, even on deep reflection, and if AP* really does require those conclusions (it might be supplemented, in a non-ad hoc way, by other principles that would save the case-specific judgment), then you are warranted in rejecting AP* or modifying it. But do not expect perfect coherence at the outset. Some of our judgments about individual cases might be mistaken.

176. *See generally* Berman, *supra* note 19; Mitchell N. Berman, *Commercial Speech and the Unconstitutional Conditions Doctrine: A Second Look at "The Greater Includes the Lesser,"* 55 VAND. L. REV. 693 (2002).

them to build a forty-foot home. They argue, among other things, that a forty-foot house on their lot would not block their inland neighbors' views and that, because nearby beachfront houses are comparably tall, their requested construction would not alter a uniform aesthetic. The commission asks the Johnsons to grant a public easement across their beach in exchange for the variance. The Johnsons refuse and the commission denies the variance.

I expect readers to share the judgments that Judge Davis's imposition of a twenty-year sentence violated Harris's constitutional rights and that the Commission's denial of the Johnsons' requested variance violated their constitutional rights. But why? After all, Harris was not constitutionally entitled to a sentence of less than twenty years, and the Johnsons were not constitutionally entitled to build a forty-foot-tall house. Constitutionally speaking, both were "benefits" that the state could withhold.

The answer, I suggest, is supplied by AP*. Harris has a constitutional right to appeal his conviction. Judge Davis penalized him for exercising this right by denying him the benefit of a lower sentence as retribution for exercise of that right. The Johnsons have a constitutional right not to have property taken from them without just compensation. That is the right they invoke when refusing to grant a public right of access across their beach. The Commission penalized them for exercising their right when denying them a benefit for the purpose of discouraging them or similarly situated others from insisting on their right in like circumstances. The combination of P* and AP explain why these actions are unconstitutional, as surely all agree that they are.[177]

In fact, this analysis corresponds extraordinarily well with actual Supreme Court decisions that approximate *Vindictive Sentencing* and *Short Zoning*: *North Carolina v. Pearce*[178] and *Nollan v. California Coastal Commission*[179] respectively.

---

177. Recall Cass Sunstein's rejection of AP. *See supra* note 154. "The clearest example" he can muster for the proposition that "the government can legitimately 'penalize' the exercise of constitutional rights through selective funding" is government's funding of public but not private schools. *See* Sunstein, *Anachronism*, *supra* note 154, at 603 & n.42, 609–10. But the non-funding of private schools, even when conjoined to the funding of public schools, does not, according to P*, penalize parents' right (grounded in the Free Exercise Clause) to send their children to private school. That non-funding of private schools makes it harder or more costly to exercise parents' rights over the education of their children need not figure into the government's reasoning at all. Because public and private schooling may differ in various ways—including regarding the extent to which each tends to promote class and racial integration and the extent to which government can influence the curriculum—the state may on legitimate grounds value the former more highly than the latter. Put differently, the state may decide that free education open to all members of the community, provided and shaped by the polity in a collective capacity, is a distinct type of good, and one worth providing.

178. 395 U.S. 711 (1969).

179. 483 U.S. 825 (1987).

Pearce involved consolidated challenges to longer criminal sentences imposed after defendants successfully appealed a first conviction but were convicted again after retrial.[180] In contrast to the hypothetical *Vindictive Sentencing* however, in neither case did the resentencing judge announce his reasons for the longer sentence.[181] The Court started by declaring basic principles:

> It can hardly be doubted that it would be a flagrant violation of the Fourteenth Amendment for a state trial court to follow an announced practice of imposing a heavier sentence upon every reconvicted defendant for the explicit purpose of punishing the defendant for his having succeeded in getting his original conviction set aside. Where, as in each of the cases before us, the original conviction has been set aside because of a constitutional error, the imposition of such a punishment, "penalizing those who choose to exercise" constitutional rights, "would be patently unconstitutional." And the very threat inherent in the existence of such a punitive policy would, with respect to those still in prison, serve to "chill the exercise of basic constitutional rights." But even if the first conviction has been set aside for nonconstitutional error, the imposition of a penalty upon the defendant for having successfully pursued a statutory right of appeal or collateral remedy would be no less a violation of due process of law. "A new sentence, with enhanced punishment, based upon such a reason, would be a flagrant violation of the rights of the defendant."[182]

In short, "Due process of law, then, requires that vindictiveness against a defendant for having successfully attacked his first conviction must play no part in the sentence he receives after a new trial."[183]

Because "[t]he existence of a retaliatory motivation would, of course, be extremely difficult to prove in any individual case,"[184] the majority proceeded to announce that vindictiveness would be conclusively presumed "whenever a judge imposes a more severe sentence upon a defendant after a new trial," unless the reasons for the more severe sentence "affirmatively appear."[185] Twenty years after *Pearce*, in *Alabama v. Smith*,[186] the Court overruled this strict prophylactic rule.[187] But no Justice in the *Pearce* line of cases has disputed the general principle that a vindictive reason for giving a criminal defendant a longer sentence than he had received previously is unconstitutional even where the defendant is not constitutionally entitled to a

180. *Pearce*, 395 U.S. at 713–14.

181. *Id.* at 726.

182. *Id.* at 723–24 (internal citations omitted).

183. *Id.* at 725.

184. *Id.* at 725 n.20.

185. *Id.* at 726.

186. 490 U.S. 794 (1989).

187. *See id.* at 795 ("We hold that no presumption of vindictiveness arises when the first sentence was based upon a guilty plea, and the second sentence follows a trial.").

shorter sentence.[188] Moreover, no Justice has taken issue with the *Pearce* majority's observation that the vindictive sentence is unconstitutional because it amounts to a forbidden "penalty."[189]

The basic idea applies to vindictiveness outside of the criminal justice context. In *Perry v. Sindermann*,[190] for example, a teacher in the Texas state college system alleged that the state declined to renew his contract because, as president of the local teachers association, he had criticized the Board of Regents.[191] In reasoning and language that nearly mirrors the general principle endorsed by *Pearce*, the Supreme Court reiterated that,

> even though a person has no "right" to a valuable governmental benefit, and even though the government may deny him the benefit for any number of reasons, there are some reasons upon which the government may not rely. It may not deny a benefit to a person on a basis that infringes his constitutionally protected interests—especially his interest in freedom of speech. For if the government could deny a benefit to a person because of his constitutionally protected speech or associations, his exercise of those freedoms would in effect be penalized and inhibited.[192]

Perry's declaration that "there are some reasons upon which the government may not rely," in particular that it "may not . . . deny a benefit to a person because of his constitutionally protected speech or associations" restates what I described as the core idea behind P*: "[T]he state may not take the expected fact that a proposed course of action would make the exercise of rights more costly or more painful as a reason in favor of that course of action."[193] Penalty skeptics have not given this possibility a serious hearing.

Nollan is much like *Short Zoning*, except that it lacks a record in which relevant governmental actors announce that they impose more stringent zoning rules than they believe are necessary to serve the public interest, for the purpose of using the offer of a variance to extract a waiver of rights that the state could not mandate. In the absence of that "smoking gun," the question in *Nollan* became whether such a purpose could be inferred from the fact that the zoning rule and the extraction demanded as a condition for its non-enforcement served somewhat different purposes: the height limitation served the public's interest in being able to see the coast from

188. *See id.* at 798–99 (surveying and approving of cases that used the presumption of vindictiveness).

189. *Pearce*, 395 U.S. at 724.

190. 408 U.S. 593 (1972).

191. *Id.* at 594–95.

192. *Id.* at 597.

193. *See supra* note 168 and accompanying text.

some distance inland; the easement served the public's interest in being able to traverse the beach.[194]

The Supreme Court divided, 5–4, on just this question. A majority thought the purposes that would constitute a penalty—in the sense marked out by P*—could be inferred from the structure of the proposal: "unless the permit condition serves the same governmental purpose as the development ban, the building restriction is not a valid regulation of land use but an out-and-out plan of extortion."[195] The dissenters thought the inference unwarranted.[196] But they did not disagree that, if denial of the requested permit were in fact animated by the purposes that the majority ascribed to the land use commission, the denial would unconstitutionally penalize the Nollans' Fifth Amendment rights.[197]

Notice this. "Extortion" is fairly understood as theft by coercion. In the majority's estimation, then, the Commission's offer to the Nollans—we'll give you a construction permit if and only if you cede a lateral easement to the public—violated an anti-coercion principle.[198] But that anti-coercion principle is manifestly not an anti-compulsion principle. Suppose the Nollans wanted only to make a modest addition to their existing home. The threat to deny permission to do so would not, in that event, give them "no choice" or even "no practical choice" other than to accept: they could live happily as they were. Still, the threat would impermissibly threaten a penalty, on the majority's estimation. Remarkably, *Nollan* was decided just three days after *Dole*.[199] Both cases raised the conditional offer problem, and both evaluated the conditional offers before them against principles fairly described in "anti-coercion" terms. But the *Dole* "anti-coercion principle" is really an anti-compulsion principle, while *Nollan* endorsed an anti-coercion principle. *Nollan* was on sounder normative footing.

---

194. Nollan v. Cal. Coastal Comm., 483 U.S. 825, 836–37 (1987).

195. *Id.* at 837 (internal quotation marks and citations omitted).

196. *Id.* at 849–50. In my view, the dissenters had the better of the argument. The majority's assertion that "the lack of nexus between the condition and the original purpose of the building restriction converts that purpose to something other than what it was," *id.* at 837, is nonsense. Nonetheless, the majority's approach might make sense if understood, not as a claim about metaphysics or logical deduction, but instead as a determination that the judiciary should police exactions by means of a judicial rule that conclusively presumes a conditional permit offer to threaten a penalty when the public interests served by the restriction and by the exaction differ. This, however, would be a prophylactic rule—what I would term a prophylactic "decision rule." (On the meaning of "constitutional decision rules," see *infra* note 231 and accompanying text.) That would be fine with me, but uncomfortable for Justice Scalia, the author of *Nollan*, given his jeremiad against prophylactic rules in his *Dickerson* dissent. Dickerson v. United States, 530 U.S. 428, 457–61 (2000) (Scalia, J., dissenting).

197. *See Nollan*, 483 U.S. at 843 (Brennan, J., dissenting) (suggesting that if the majority were correct that the regulation exceeded the state's police power, it would be an unconstitutional taking).

198. *See, e.g.*, BLACK'S LAW DICTIONARY 623 (8th ed. 2004) (defining "extortion" as "obtaining something . . . by illegal means, as by force or coercion").

199. *Nollan*, 483 U.S. 825 (1987) (decided on June 26th); South Dakota v. Dole, 483 U.S. 203 (1987) (decided on June 23rd).

As *Pearce* and *Nollan* illustrate, it will often be hard to determine whether a given non-provision of a benefit *is* a penalty in the sense captured by P*, and also, therefore, whether some conditional offer of a benefit threatens a penalty in that same sense. More fundamentally, though, they teach that, epistemic difficulties aside, the state may not penalize rights. Without qualification or dissent, they affirm AP*.

4. Beyond the Hypothetical.—The previous subsection aimed to bolster the plausibility of AP* by analyzing hypothetical cases in which the types of reasons or purposes necessary to make out a penalty were patent. The actual cases that I paired with the hypotheticals raise the question of whether we can ever infer the bad purposes from the structure of the proposal itself, without having to put words into the mouths of the key governmental actors. I pursue that question here—and answer it in the affirmative—by analyzing the most important conditional spending precedent: *South Dakota v. Dole.*

Dole involved a challenge to federal highway spending law that conditioned 5% of the funds that a state would be authorized to receive for highway construction and repair on its maintenance of a minimum legal drinking age (MLDA) of at least 21.[200] The Court, recall, determined that the proposal was not unconstitutionally "coercive," but interpreted "coercion" to mean compulsion:

> When we consider, for a moment, that all South Dakota would lose if she adheres to her chosen course as to a suitable minimum drinking age is 5% of the funds otherwise obtainable under specified highway grant programs, the argument as to coercion is shown to be more rhetoric than fact. . . . Here Congress has offered relatively mild encouragement to the States to enact higher minimum drinking ages than they would otherwise choose. But the enactment of such laws remains the prerogative of the States not merely in theory but in fact.[201]

The Court was surely right to conclude that the proposal did not constitute compulsion. I have argued, however, that the normatively meaningful concept is coercion. And the offer would have been coercive if it would have been unconstitutional for Congress to withhold the offered benefit (some portion of federal highway funds) on failure of the condition. Furthermore, given the specification of the anti-penalty provision captured by AP*, non-provision of that benefit would have been unconstitutional had it been motivated by a purpose to discourage or punish exercise of a state's right to maintain a MLDA under 21.

Keep in mind: we are inquiring into the reasons the offeror would have for *withholding the benefit* at t_2 on non-satisfaction of the stated condition; we are *not* inquiring into the public-serving reasons for *extending the*

200. 483 U.S. at 203.

201. *Id.* at 211–12.

proposal, or for attaching this particular condition, at t_1. (This is an absolutely critical distinction that even sophisticated readers have missed;[202] if you glide past it, then you have no chance to understand the analysis.) Often, though not invariably, government is constitutionally permitted to offer benefits on condition precisely as a means of inducing a rightholder to waive the protection of a constitutional right or to exercise her right in a manner that the government prefers.[203] The conditional spending power, plea bargains, and even governmental employment all depend upon this fact. It is crucial, however, that in many or most cases where the deployment of such conditional offers is permitted, government would lack (much or any) affirmative reason to provide the offered benefit if the offeree refuses to abide by the stated condition, in which event the failure, at t_2, to provide the offered benefit would not itself be motivated, as was the offer at t_1, by a waiver-inducing purpose. This is true of prosaic offers not involving the government too: if I offer you $10 for the shirt off your back, and if you decline, then what best explains my consequent failure to provide you with the benefit of $10 is simply that I lack an affirmative reason to provide it and not that I have affirmative reason to provide it but allow that affirmative reason to be overridden by a punitive or waiver-inducing purpose.[204]

There's a pretty simple test for determining whether the offeror would have acceptable reasons for withholding the benefit. This test is imperfect but good as a first pass. Imagine two things: first, that there is only a single offeree, not a class of them; and second, that the offeror knows that the offeree will not accept the deal, i.e., that it will not comply with the condition. Would the offeror, if genuinely motivated to advance the public interest, nonetheless withhold the benefit at issue? If so, then the withholding of the benefit does not penalize the offeree for exercising its right. If not, then the withholding of the benefit does penalize the offeree for exercising its right, in which case the conditional proposal threatens an unconstitutional penalty, hence constitutes the constitutional wrong of coercion.

A hypothetical contrasting case facilitates the analysis.[205] Suppose that by 1984 every state had a minimum drinking age of 21, except for South Dakota, which maintained a legal drinking age of 18, and that each state had a minimum *driving* age of 18, except for North Dakota, which imposed a minimum driving age of 55. Wanting to induce each state to change its outlying policies, Congress

---

202. *See, e.g.*, Bagenstos, *supra* note 167, at 378 (erroneously stating that "Berman treats a federal funding condition as imposing a penalty whenever *the law* has the purpose of influencing the states' behavior") (emphasis added) (internal quotation marks omitted).

203. When this is, and is not, a permissible purpose for governmental conduct must be determined by provision- or rights-specific analysis; it cannot be resolved by the principles or considerations that are general to a trans-substantive framework for thinking through the conditional offer problem.

204. This is a profoundly important distinction. We will return to it.

205. This comparison first appeared in Baker & Berman, *supra* note 51, at 537–39.

directed that a state would lose 5% of the highway funds it would otherwise receive if it maintained a minimum drinking age under 21 and would lose all of its highway funds for imposing a minimum driving age over 18. Both conditions amount to congressional threats to withhold a benefit.[206]

But as I have just emphasized, that alone can't make either proposal coercive. On our best account of coercion, the proposal is coercive if carrying out the threat would be unconstitutional, and, per AP*, carrying out the threat would unconstitutionally penalize the states' (presumed) sovereign right to set a drinking or driving age as it wishes if done in order to make the exercise of such a right costlier.

Now imagine, though, that both Dakotas reject the conditions. What interests would justify Congress in carrying out its threat to withhold highway funds—5% of South Dakota's, all of North Dakota's? The story with respect to North Dakota might go like this. The higher the minimum driving age, the smaller the number of cars on the roads and the smaller the number of accidents. If the latter numbers are very small, then improvements to road conditions could net only a very small reduction in accidents and their associated costs. Therefore, every federal dollar spent on North Dakota road improvements purchases a much smaller social welfare benefit than is purchased in the other states. So if North Dakota (or any other state) insists on maintaining an unusually high minimum driving age, federal funds could produce a higher return in their next best use than when devoted to highway improvements in that state.[207]

It's all well and good for a state to maintain a very high driving age, Washington might therefore think, but because the highways in such states will be so underused, the national interest is not well-served by improving them. In short, withholding the offered funds on failure of the driving-age condition need not serve any interest in punishing North Dakota or in shaping state behavior, which is to say that withholding the funds does not penalize North Dakota, in which event the conditional threat to withhold such funds is not coercive.[208]

This story is rather less plausible with respect to South Dakota, however.[209] To be sure, improving road conditions and raising the minimum

206. *Id.* at 537.

207. *Id.*

208. *Id.*

209. In the real world, of course, this story is not very plausible with respect to North Dakota either. For one thing, Congress could (in fact, does) introduce annual highway miles driven into the ordinary formula for allocating highway funds, in which case introducing driving age as a separate factor would be redundant. But this driving age hypothetical is designed merely to show that not all conditional spending proposals involve threats to withhold federal funds under circumstances in which such withholding would be undertaken for an improper reason. It illustrates that proposition by showing what form a counterexample would take even if it would not itself, in all probability, constitute such a counterexample. In any event, any objection to the example could be met by tweaking the hypothetical. So, for example, I could ask you to imagine that the technology necessary to measure annual highway miles driven does not exist or is prohibitively expensive to employ.

drinking age (from 18 to 21) might each increase net social welfare. But that's not the issue. The issue is whether the extent to which improving road conditions increases net social welfare is itself contingent upon the minimum drinking age. Put another way, the issue is whether the *increase* in highway safety that Congress would buy by giving a state funds with which to improve its highways varies depending upon that state's minimum drinking age, such that the higher a state's MLDA (within the relevant range), the greater is the increase in highway safety that federal highway dollars purchase. Because if it doesn't, then withholding federal highway funds on failure of the condition does not serve a legitimate federal interest except as mediated by a purpose—the type of purpose that turns a permissible non-provision of a benefit into an impermissible imposition of a penalty—to discourage states from refusing the federal demand. That is, if $X spent on highway maintenance and construction would reduce highway accidents (or injuries or accident costs) *y* regardless of whether the state has an MLDA of 18 or 21 (albeit from different baselines), then Congress's non-provision of some portion of that $X because a state maintains the lower MLDA is only intelligible as a means to punish the recalcitrant state or to discourage other states from similarly refusing the federal condition.

All of this is put conditionally. So, what are the facts? Are road improvements less valuable in states with lower drinking ages, all else equal? It is hard to imagine why they would be. If anything, it is more plausible to suppose that road improvements buy marginally *greater* decreases in accidents where driving conditions are more dangerous, such as where a greater percentage of drivers are impaired. In any event, nothing in *Dole* or the relevant legislative history suggests even remotely that any member of the Court or of Congress believed road improvements are of less value in states with lower drinking ages. The conclusion is thus warranted, if not quite inescapable, that withholding highway funds from South Dakota served a purpose in punishing or discouraging the exercise of a state's right to set its own MLDA. The action that Congress's offer threatened would therefore violate the anti-penalty principle, and the threat itself would thus violate the anti-coercion principle. *Dole* was wrongly decided.

While I do not find this conclusion jarring, I know from conversation that some constitutional scholars find the correctness of *Dole* much harder to give up. I would simply urge readers who share that view to reconsider. As a spur to reconsideration, the reader whose sympathies run more liberal and more nationalist than do those of most critics of either *Dole* or the Affordable Care Act might reflect on two hypotheticals: (1) Congress conditions some (significant) percentage of federal funds for education on a state's continued criminalization of marijuana use, either generally or for medical purposes; and (2) Congress conditions some (significant) percentage of federal funds for economic development projects on a state's elimination or non-adoption of laws that prohibit discrimination in the private sector on the basis of sexual orientation. It is reasonably clear that both laws would pass muster

under *Dole*.[210] Granted, this is not an *argument* against *Dole*. But some readers may find that it pumps anti-*Dole* intuitions.

More generally, our task, as I see it, is to distill general constitutional principles that seem plausible on their own and that best explain and justify a large set of intuitions that survive reflection about the proper resolution of a large set of conditional offer cases. Our goal should be to articulate and refine a set of general principles that best cohere with case law, with intuitively sensible outcomes across the unconstitutional conditions space, and with yet more general normative principles that seem plausible and attractive and that have explanatory power in their domains, all while keeping in mind that the principles that cohere "best" might still cohere imperfectly. As in any exercise designed to achieve reflective equilibrium, we must be prepared to give up some intuitions with which we start. With that in mind, it's not as though analysis that begins with a philosophically defensible interpretation of the anti-coercion principle and with a conception of penalty that vindicates an anti-penalty principle yields conclusions that undermine *McCulloch* or *Marbury* or *Brown*. Let us not treat *Dole* as sacrosanct. It can be abandoned.[211]

B. The Three-Offer Analysis

We now have most of the tools necessary to determine whether the Medicaid expansion is unconstitutionally coercive on the grounds that it threatens to penalize the exercise of the state's constitutional rights. I propose to address that question in two steps. In this section, I analyze the proposal as the Chief Justice did, namely as a new program distinct from the rest of Medicaid. In the next, I investigate whether this is the best or fairest way to parse the Medicaid expansion and, if not, what should be the constitutional bottom line.

What does it mean to view the Medicaid expansion as a new program, distinct from the rest of Medicaid? It means, I think, that the entire bundle is

---

210. These hypothetical statutes would easily satisfy four of the *Dole* requirements, at least under present doctrine: they are for the general welfare and unambiguous; they do not violate any independent constitutional bar; and they do not pass the point at which pressure becomes compulsion. They would also pass the germaneness prong so long as Congress could identify a purpose for the condition that "bear[s] some relationship to the purpose of the federal spending." New York v. United States, 505 U.S. 144, 167 (1992). That would be child's play. (1) Decriminalization of marijuana is objectionable in part because it is likely to increase marijuana use by minors, thus impeding their intellectual development; the condition and the funding are both geared toward improving children's intellectual development. (2) Private anti-discrimination laws are objectionable (even if justifiable, all things considered) in part because they cause risk-averse private actors to take economically inefficient precautions against liability; the condition and the funding are both geared toward promoting economic growth. For elaboration on the ease with which *Dole*'s relatedness prong can be satisfied, see Baker & Berman, *supra* note 51, at 499–503.

211. But what if *Dole* was rightly decided? How far, wonders Glenn Cohen in private correspondence, can readers who would not abandon *Dole* despite my analysis follow me? It's impossible to say. It all depends upon the particular reasons one has for finding my analysis of *Dole* unpersuasive.

properly conceptualized as consisting of three biconditional proposals. In simplified and stylized form, they are as follows:

Proposal 1 (the preexisting Medicaid program): We (the federal government) will give you (a state) $X for the medical needs of the blind, the disabled, the elderly, and needy families with dependent children in your state if and only if you comply with various conditions, C1 (that we are disabled from mandating).

Proposal 2 (the new Medicaid expansion): We (the federal government) will give you (a state) $Y for the medical needs of the childless poor adults if and only if you comply with various conditions, C2 (that we are disabled from mandating).

Proposal 3 (an ACA requirement): We (the federal government) will make you (a state) eligible to receive and thus to accept Proposal 1 if and only if you accept Proposal 2.

To contend, as the state challengers did, that the Medicaid expansion is unconstitutional because it threatens to withdraw all Medicaid funding from states that do not agree to the conditions on receipt of funds for a new class of beneficiaries is just to contend that Proposal 3 is unconstitutional. Given AP*, the constitutionality of Proposal 3 depends on the reasons the federal government would have for withdrawing a state's eligibility to accept Proposal 1 in the event that it does not accept Proposal 2.[212] In particular, Proposal 3 would unconstitutionally penalize a state's supposed constitutional right to decline Proposal 2[213] if carrying out the action threatened would be animated by a purpose in making the exercise of that right more costly or painful. (Hence solutions to the second and third puzzling features of Chief Justice Roberts' opinion—what I termed the Reasons Riddle and the Penalty Puzzle: whether the Medicaid expansion abridged the anti-coercion principle depends upon whether it threatened to penalize the states' rights, and whether the act threatened would amount to a penalty depends upon the reasons or purposes fairly ascribable to Congress.)

Surely Congress would have the proscribed purposes were it to carry out the act that Proposal 3 (call this "the Linking Proposal") threatens.[214] Even if not constitutionally obligated to do so, Congress has good and

---

212. "Depends" is a little too strong. Conceivably, Proposal 3 could be unconstitutional even if it does not threaten a penalty and hence isn't coercive. *See supra* note 19. In this case, though, no other basis for its unconstitutionality seems remotely likely.

213. *See supra* note 20.

214. "Wired" plea bargains in which a plea bargain offered one defendant is conditioned not only on her pleading guilty, but also on her co-defendant accepting a plea bargain offered him, can also be analyzed as two separate conditional offers supplemented by a linking proposal. Whether the linking proposal in wired plea bargains is unconstitutionally coercive for threatening to penalize a defendant's constitutional right to put the state to its burden of proof is a separate question that I do not address here, except to register my disagreement with the D.C. Circuit's observation that the answer to this question reduces to "whether the practice of plea wiring is so coercive as to risk inducing false guilty pleas." U.S. v. Pollard, 959 F.2d 1011, 1021 (D.C. Cir. 1992). I am grateful to Dan Markel for drawing the practice, and the case, to my attention.

legitimate reasons for granting a state funds to provide for the medical needs of disabled persons, blind persons, and poor families with children. These reasons are essentially ones of humanity or beneficence.[215] Now suppose that some state—Florida, let's say—chooses not to comply with the conditions (C2) necessary to receive additional federal funds earmarked for the medical needs of poor, childless adults. It is hard to imagine how that fact cancels or weakens the reasons Congress has to provide the funds described in Proposal 1. The blind, disabled, and poor children in Florida are just as in need of public medical assistance and just as deserving (however needy or deserving that might be), regardless of Florida's decision with respect to Proposal 2. Therefore, the conclusion is nearly irresistible that the federal government's purpose (or, if you prefer, a purpose fairly attributable to the federal government) for withholding the benefit of eligibility for Proposal 1 on failure of Florida to comply with the condition in Proposal 3 is to make it costly for Florida to exercise its constitutional right to decline Proposal 2, thereby inducing it to change its decision or discouraging other states from following Florida's example. On our best rendering of the anti-penalty principle—the rendering captured by AP*—that is simply not a permissible reason for the government to treat a rightholder less well than it otherwise would.

If the analysis in the preceding paragraph is correct, and if the Medicaid expansion is fairly viewed as a new and distinct program, we are almost ready to conclude that the Medicaid expansion is unconstitutionally coercive for unconstitutionally threatening to penalize a state's exercise of its constitutional rights. Almost, but not fully. I said two paragraphs ago that, given AP*, Proposal 3 would unconstitutionally penalize a state's supposed constitutional right to decline Proposal 2 if carrying out the action threatened would be animated by a purpose in making the exercise of that right more costly or painful. That is not exactly what AP* says. The anti-penalty principle speaks in terms of "presumptive" unconstitutionality. That is, penalizing rights is pro tanto or defeasibly unconstitutional, but potentially justifiable. In a familiar vocabulary, to penalize a right is to *infringe* the right but not necessarily to *violate* it. It is therefore open to defenders of the ACA to argue that even if carrying out the threat contained in Proposal 3 penalizes a state's constitutional rights, doing so is justified by the national

215. I reiterate that it is essential to distinguish two situations in which the national government does not provide an offered benefit after a state executes its constitutionally protected decision not to comply with a stated condition. *See supra* note 204 and accompanying text. In the first, the failure of the condition leaves the national government without affirmative reason to provide the benefit; in the second, the national government has affirmative reason(s) to provide the benefit notwithstanding failure of the stated condition but allows such reason(s) to be overridden by countervailing reasons. In the first type of case, withholding the benefit will not penalize exercise of the right. In the second, withholding the benefit will penalize exercise of the right if the overriding reason involves a purpose to make the state's exercise of its constitutionally protected decision more costly.

government's weighty interests in improving the provision of health care in this country by making it more effective and less expensive.

The extent to which (a) the threat to penalize a state's constitutional prerogatives would in fact advance this national interest and (b) this national interest could not be advanced comparably well by means that do not call forth a demand for heightened justification are matters that depend upon messy empirical assumptions and causal hypotheses; they cannot be thoroughly evaluated from the comfort of a constitutional theorist's armchair. Therefore, I will content myself with two observations. First, the dispute would no longer be about whether the Medicaid expansion is coercive in a constitutionally meaningful sense—by hypothesis, it is—but about whether it is *justifiably* coercive. Second, however we should assess whether the justificatory burden is satisfied, whether by the compelling-interest test or otherwise, it cannot be enough that the Medicaid expansion serves valuable ends. The whole point of anti-penalty and anti-coercion principles is that constitutional rights impose significant constraints on the means that the state may adopt even in pursuit of good goals. I am highly skeptical that this coercion can be justified, but acknowledge that the question should be considered open—though, in my view, only ajar.

C. *A Package Deal—Or Not?*

I think it fairly plain that the Medicaid expansion at least infringes a true anti-coercion principle when conceived as the conjunction of three conditional proposals. This explains why Chief Justice Roberts took pains to describe the Medicaid expansion as a new and distinct program.[216] This section addresses whether his conclusion really did depend, as he seemed to believe it did, upon his contested characterization of the Medicaid expansion. We can break this fundamental question into two subordinate ones: First, assuming arguendo that the Medicaid expansion is unconstitutionally coercive if fairly viewed as a new program, is it also unconstitutionally coercive if fairly viewed only as a modification of, or amendment to, the existing Medicaid program? Second, if not, how do we adjudicate the dispute between the majority and Justice Ginsburg regarding how the Medicaid expansion is "properly viewed"?[217] (Notice that, no matter our answers to these two questions, we have a good solution to the first of the three puzzles identified in Part III—the Modification Mystery. Whether the Medicaid expansion is separate from the rest of Medicaid seemed clearly irrelevant on anti-compulsion reasoning. It looks likely to be relevant on anti-coercion reasoning even if careful analysis persuades us that it isn't.)

---

216. *See* Nat'l Fed'n of Indep. Bus. v. Sebelius, 132 S. Ct. 2566, 2575 (2012) (Roberts, C.J.) (stating that "the expansion accomplishes a shift in kind, not merely degree").

217. *Id.* at 2605.

Although this is a natural way to proceed, I think it will turn out not to be felicitous. There is no metaphysical truth regarding whether some set of benefits offered on some set of conditions is one program or a combination of programs each constituted by some subset of all benefits offered on some subset of all the conditions.[218] My instinct is to formulate the question in normative rather than metaphysical terms. In particular, we should ask whether, even allowing Congress to designate any bundle of offers as a single program, a state challenger should be entitled to insist that courts analyze the program as smaller conditional offers, in which acceptance of one serves as an additional condition for another, on the model employed in the previous section. We can call this the disaggregation problem.

A solution to the problem starts by acknowledging that neither polar position is tenable. On the one hand, an offeree cannot have carte blanche to carve programs as it sees fit. Consider the employment context. Simplified, the deal proposed by a state employer to a would-be employee is: "If you agree to conditions a, b, c, d, e, and f, we agree to give you $X." If the employee were permitted to disaggregate this bundled offer into separate conditional offers, we'd be forced to allocate percentages of $X to each condition, and I see no good way to do that. Medicaid itself (even putting the ACA expansion aside) is an extraordinarily complex program that could be parsed as a bundle of hundreds or thousands of analytically distinct conditional offers. On the other hand, the governmental offeror does not have unlimited freedom to bundle discrete deals into one massive deal. Surely Congress could not lump all its present conditional spending deals (for education, highways, Medicaid, etc.) into a single "Super Program" that offered a huge sum in exchange for compliance with a vastly large set of conditions. To allow that gambit would be to eviscerate any in-principle limit on the federal government's ability to manipulate states into doing its bidding. Perhaps that would be a better system, but it is disingenuous to contend that such a system would be faithful to the interests or values that underlie our system of federalism. Unfortunately, while neither extreme position is acceptable, no test or standard for navigating between the poles presents itself as obvious.[219] I am disposed to believe that the disaggregation problem is genuinely hard.[220]

---

218. Compare an observation made earlier: instead of asking, following the lead of Robert Nozick, whether some biconditional proposal itself is a threat or an offer, we should ask whether the conditional threat that is one component of the proposal is wrongfully coercive. *See supra* note 55 and accompanying text.

219. "Germaneness" or "relatedness" reasoning won't do the trick. *See* Baker & Berman, *supra* note 51, at 512–17.

220. For other recognition of both the importance and difficulty of the problem, see Richard H. Pildes, *Avoiding Balancing: The Role of Exclusionary Reasons in Constitutional Law*, 45 HASTINGS L.J. 711, 736–41 (1994).

A preliminary step toward proposing a solution is to identify the potentially relevant factors or considerations.[221] Here are several: (1) whether the provisions that constitute the putative single program were adopted all at once or separately; (2) the extent to which the type and amount of benefit can be allocated to distinct conditions or groups of conditions objectively or, instead, would be arbitrary or require inescapably contestable judgments; (3) the extent to which realization of the purposes behind one disaggregated conditional offer depends upon satisfaction of a separate disaggregated conditional offer; and (4) the extent to which allowing the offerees to pick and choose among conditions would burden administration of the program. If these and other candidate factors point in the same direction with respect to any specific proposal to disaggregate what the offeror would present as a single program, then courts may provisionally resolve that particular dispute while deferring to a later case the more difficult work of determining just which of these factors are relevant and just how they should be combined—in a multi-factored balancing test or in something more rule-like.

In the case of the Medicaid expansion, all four of these factors seemingly do point in the same direction—in support of disaggregation. (1) The Medicaid expansion was enacted after a coherent program (itself the product of many statutes over many years) already existed.[222] (2) It clearly identifies the new conditions that must be satisfied to receive new dollars.[223] (3) The medical needs of each class of beneficiaries can be served whether or not a state agrees to serve the needs of other classes.[224] (4) Allowing states to opt out of the expansion would not appear to create substantial administrative difficulties for the Department of Health and Human Services.[225]

If I am correct that each factor by itself weighs in favor of disaggregation, then states should be entitled to have courts analyze the

---

221. For the moment, I'll pass over whether a given factor is relevant causally or constitutively, on the one hand, or merely evidentiarily, on the other.

222. *See Medicaid: A Timeline of Key Developments*, KAISER FAMILY FOUND., http://www.kff.org/medicaid/timeline/pf_entire.htm (describing the various statutes that have impacted the Medicaid program over the years).

223. *See* KATHLEEN S. SWENDIMAN & EVELYNE P. BAUMRUCKER, CONG. RESEARCH SERV., SELECTED ISSUES RELATED TO THE EFFECT OF *NFIB V. SEBELIUS* ON THE MEDICAID EXPANSION REQUIREMENTS IN SECTION 2001 OF THE AFFORDABLE CARE ACT 3 (2012), *available at* http://www.ncsl.org/documents/health/aca_medicaid_expansion_memo_1.pdf (stating that the allocation of federal funds depends upon whether states meet the ordinary Medicaid standards or meet the higher standards established by the ACA).

224. *Id.*

225. The Congressional Budget Office's estimates, updated after the Health Care Decision, do not address administrative costs, indicating that those costs are not substantial. *See* CONG. BUDGET OFFICE, ESTIMATES FOR THE INSURANCE COVERAGE PROVISIONS OF THE AFFORDABLE CARE ACT UPDATED FOR THE RECENT SUPREME COURT DECISION 5 n.9 (2012), *available at* http://www.cbo.gov/sites/default/files/cbofiles/attachments/43472-07-24-2012-CoverageEstimates.pdf (the updated estimates "do not include federal discretionary administrative costs, which will be subject to future appropriation action").

conditional offers as discussed in the previous section even if Justice Ginsburg gets the better of Chief Justice Roberts in their debate over whether the Medicaid expansion "is in reality a new program"[226] (and assuming that that is a meaningful question). Tentatively and provisionally, then, courts should analyze the Medicaid expansion as the trio of offers described in the previous section, and should therefore accept the conclusion already advanced: the ACA threatens to penalize the states' right to decline to provide health coverage for a new class of beneficiaries, and thus runs afoul of the normatively meaningful "anti-coercion principle."

There is another possible way to resolve the disaggregation problem that similarly avoids the need for courts to resolve whether some cluster of benefits and conditions is "in reality" one program or more, but is more structured, less impressionistic. At the first stage of analysis, courts should allow an offeree to disaggregate a putative program into distinct conditional offers in whatever fashion it chooses so long as it provides persuasive grounds for linking the benefits and demands as it does. Imagine a program that offers benefits {B1, B2, . . . Bn} to states that agree to conditions {C1, C2, . . . Cn}. If the state offeree is willing to comply with all conditions except C2 and proposes to decouple the conditional offer of benefit B1 on condition C2, in order to comply with the complex conditional offer that remains, it must explain why C2 pairs with B1 and not with, for example, B2. This is essentially to treat factor (2) as a threshold requirement.

If the offeree can pass this threshold, then the second stage of analysis directs courts to evaluate the program in disaggregated form. In particular, it directs them to determine whether "the Linking Proposal" is coercive—a question that, I have argued, depends on the reasons the offeror (the federal government in cases of conditional offers to the states) would have for carrying out the threat to deny eligibility for the conditional offer that remains after decoupling. It is at this second stage that factors (3) and (4) become relevant. If a state's noncompliance with condition C2 either would frustrate the interests that compliance with conditions *except* for C2 would otherwise serve (i.e., if complementarity among conditions obtains), or would create significant administrative difficulties, then it is not the case that the offeror, in carrying out the Linking Proposal threat to withhold benefits, would act for the purpose of making it costly for states to exercise their supposed rights to decline condition C2. It strikes me as reasonably clear that the Medicaid expansion would not survive this more structured analysis.

---

226. Nat'l Fed'n of Indep. Bus. v. Sebelius, 132 S. Ct. 2566, 2605 (Roberts, C.J.). *Compare id.* at 2605–06 (Roberts, C.J.) *with id.* at 2635–36 (Ginsburg, J., dissenting in part) (illustrating the difference between Roberts's view of the expansion as a new program and Ginsburg's opposing view). Roberts's characterization of the Medicaid expansion and its relationship to the history of amendments to the Medicaid program is powerfully criticized in Nicole Huberfeld et al., *Plunging into Endless Difficulties: Medicaid and Coercion in* National Federation of Independent Business v. Sebelius, 93 B.U. L. REV. (forthcoming 2013).

In any event, given the centrality of reasons to my analysis, the most fundamental point can be simply encapsulated: "We insist on the Linking Proposal because that is what the program requires" is not an adequate response by the federal government to a state requesting disaggregation. In our ordinary lives, we treat that as a "bureaucratic" answer, in the pejorative sense, and rightly reject it with exasperation. We should reject it in this context too.

V. Frequently Advanced Challenges (FACs)

In this final Part, I raise and respond to the objections to my analysis that I have encountered most often. Some of these objections are simply mistaken. Others helpfully invite clarification or qualification that I have reserved for this stage.

Objection 1: "Your analysis depends on the assumption that the constitutionality of state action can depend upon the reasons or purposes for which a legislature acts. But the Constitution does not police purposes."

Response: Oh, please. Of course it does, as many commentators have repeatedly and persuasively shown.[227] The best way to read most decisions that state or suggest otherwise is as declaring not that the constitutionality of legislative or executive action cannot depend upon the reasons, purposes, or motives that lie behind the challenged action,[228] but rather that courts ought not to inquire into those reasons, purposes, or motives. (This is sometimes clear enough from the opinion itself, but sometimes requires a little charity in interpretation.)

The distinction lies at the heart of what I have elsewhere dubbed the "two-output thesis."[229] On this picture of the logic of constitutional adjudication, courts do two things in constitutional adjudication upstream from announcing a fact-specific holding: they interpret the Constitution to yield a legal norm or proposition; and they craft rules or tests—doctrine—to

227. *See, e.g.*, Caleb Nelson, *Judicial Review of Legislative Purpose*, 83 N.Y.U. L. Rev. 1784 (2008); Ashutosh Bhagwat, *Purpose Scrutiny in Constitutional Analysis*, 85 CALIF. L. REV. 297 (1997); Richard H. Fallon, Jr., *The Supreme Court 1996 Term—Foreword: Implementing the Constitution*, 111 HARV. L. REV. 56, 71–73 (1997) (analyzing constitutional inquiries into legislative purposes).

228. I have been speaking of actions and reasons (and kindred notions) as distinct things: there is an action of not providing a benefit and there are reasons, purposes, or motives for which an actor (here, Congress or "the national government") might engage in that action. But it is also possible to inscribe reasons or purposes within the actions themselves, in which case we could isolate the action of (for example) not providing a benefit for the purpose of making the state's choice more costly. On this view, instead of asking about Congress's reasons for withholding the benefit, it would be more perspicuous to inquire into the "internal logic" of the withholding, or of the proposal. This point warrants further development; at present, I simply flag it.

229. *See* Mitchell N. Berman, *Aspirational Rights and the Two-Output Thesis*, 119 HARV. L. REV. F. 220 (2006).

implement or administer that legal norm or proposition.[230] I have called the courts' interpreted constitutional norm "a constitutional operative proposition," and the tests that courts craft and lay down for future courts to apply when determining whether the operative proposition is satisfied "constitutional decision rules."[231] But whatever the vocabulary and underlying conceptual framework, whether *courts* should police legislative or executive reasons, purposes, or motives is a separate question from whether such deliberative inputs can bear constitutively on the constitutionality of the governmental action. In general, we should think first in terms of what the Constitution, rightly interpreted, allows, commands, and prohibits. Only once we have a good handle on that, in my view, should we address what sensibly implementing judicial doctrine would look like.

Objection 2: "You rely upon a contested definition of 'penalty.' I don't think that 'penalty' is best defined as P* defines it."

Response: It is true that I believe that I have deployed an understanding of the concept of penalty that corresponds fairly well with the ordinary definition of "penalty." (The same is true with respect to coercion and "coercion.") But, as I have urged, that is not essential. Don't fixate on the words.

The substance of my claim is that it is unconstitutional to *make exercise of a right more costly than it would be but for a purpose in discouraging or punishing exercise of the right.* I then call the italicized phenomenon a "penalty." Though I believe this is an account that accords reasonably well with existing usage of the word, nothing turns on it. If you balk at that concept as a definition of our current word "penalty," fine.[232] I am once again after a concept or normative principle; I'm not playing at lexicography. That conventional meaning of the word "penalty" is of little import is reflected by the fact that AP* does not even use it.

Objection 3: "Your view denies that Congress may pursue ends through conditional spending that it could not pursue directly and thus would return

---

230. The earliest presentations of this basic view are Henry P. Monaghan, *The Supreme Court, 1974 Term—Foreword: Constitutional Common Law*, 89 HARV. L. REV. 1 (1975), and Lawrence G. Sager, *Fair Measure: The Legal Status of Underenforced Constitutional Norms*, 91 HARV. L. REV. 1212 (1978). *See also, e.g.*, Fallon, *supra* note 227; Kermit Roosevelt III, *Constitutional Calcification: How the Law Becomes What the Court Does*, 91 VA. L. REV. 1649 (2005).

231. *See generally* Mitchell N. Berman, *Constitutional Decision Rules*, 90 VA. L. REV. 1 (2004). I explore the particular use of non-standard decision rules to administer operative propositions that turn on governmental purposes in Mitchell N. Berman, Guillen *and Gullibility: Piercing the Surface of Commerce Clause Doctrine*, 89 IOWA L. REV. 1487, 1518–33 (2004) (discussing Commerce Clause doctrine), and Mitchell N. Berman, *Managing Gerrymandering*, 83 TEXAS L. REV. 781, 828–53 (2005) (discussing partisan gerrymandering).

232. Concededly, if the activity that I label penalty is too distant from ordinary usage of the word "penalty," then I am not entitled to gain support for my view from the penalty passage I quote from Chief Justice Roberts's opinion. I think that I am in fact entitled to some mileage from his passage, but I can do without it.

us to the discredited doctrine of *United States v. Butler*[233] that Congress may not use its spending power to "purchase a compliance which Congress is powerless to command."[234]

Response: No, my analysis does not revive *Butler*. Congress may try to induce behavior that it could not mandate by offering inducements just so long as it would have adequate reasons not to provide the benefits offered in the event that a state offeree declines the deal—reasons that do not depend upon the expectation that non-provision of the offered benefit would prove costly or painful to the offeree. I provided a hypothetical example in my discussion of *Dole*.[235] Proposals 1 and 2, in the disaggregated analysis of the ACA, are additional examples.

Objection 4: "Your analysis assumes that states are right holders. But it is a mistake to equate the putative 'rights' held by states with the genuine 'rights' held by individuals. Even if the Constitution is rightly interpreted to obligate government not to penalize the exercise of true rights, Congress is not similarly disabled from penalizing actions by states."

Response: My specification of the anti-penalty principle—AP*—posits that it follows from the possession of a constitutional right that the correlative duty-holder may not burden the right for certain reasons. Objection 4 can be construed to make two contentions: first, that, even if this is true of claim-rights, it is not true of those nominal rights that, in Hohfeldian terms, are privileges;[236] and second, that the "rights" that states have against the federal government are in fact privileges, not claim-rights. Whereas claim-rights correlate with duties, privileges correlate with disabilities.

I do not know what argument would support the first part of this contention. It seems to me more plausible that AP* is a corollary of privileges and of claim-rights. But perhaps "concomitant" is more apt than "corollary" here: I do not contend that AP* either is part of the concept of a right or is logically entailed by the possession of a right. So *the grounding, and therefore the scope, of AP* warrants further investigation*, leaving me open to being persuaded that the Constitution is not best understood to protect states' "rights," or some subset of them, against penalization.

Objection 5: "On your analysis, not only would the Medicaid expansion be invalid, but so too would aspects of the Medicaid program that preexisted that expansion. To see why, consider Proposal 1 in the Three-Offer Analysis. According to that Proposal, the federal government offers each state, conditioned on compliance with some specified demands, $X for the medical needs of the blind, the disabled, the elderly, and poor families with

233. 297 U.S. 1 (1936).

234. *Id.* at 70.

235. *See supra* section IV(B)(4).

236. *See generally* Wesley Newcomb Hohfeld, *Fundamental Legal Conceptions as Applied in Judicial Reasoning*, 26 YALE L.J. 710 (1917).

dependent children. But that offer could itself be disaggregated into five proposals in which each of the first four is a conditional offer of funds for one class of beneficiaries (the blind, the disabled, etc.), and the fifth is the Linking Proposal that conditions state eligibility for any one of the first four offers on a state's acceptance of the other three. Thus, if the Medicaid expansion threatens to penalize states for exercising their presumed right to decline one offer, so too did the rest of Medicaid. More generally, your analysis threatens wide swaths of federal spending programs that have not previously been suspect."[237]

Response: There is no question that the analysis I have proposed would threaten some conditional spending programs that had seemed unproblematic under *Dole*. That conclusion should not by itself prove too alarming if we can pry ourselves from the grip of the status quo bias. That said, there are several reasons why the implications of my framework for conditional spending programs are not as radical or far-reaching as might appear at first blush.

The first two I have already touched on. First, there is the disaggregation problem: many programs consisting of a bundle of conditional offers may not be disaggregable at a state's behest. Second, given the many difficulties and dangers that attend judicial inquiry into purposes, as AP* requires, courts might appropriately decide to administer these basic constitutional principles and understandings by means of under-enforcing constitutional decision rules.[238]

The third reason I have not yet emphasized, but it is more important than one might take its late appearance to signal. As I have already stressed a couple of times, the withholding of a benefit on the failure of a stated condition will not be a penalty if the failure of the condition undermines or cancels whatever reason the offeror (here, the national government) would have to provide the benefit. (If you don't agree to give me your shirt, I simply lack reason to give you the $10 I had offered.) That the national government would have some affirmative reason to provide the benefit notwithstanding a state's decision not to comply with a stated condition is thus a necessary condition for the non-provision of the benefit to constitute a penalty. What I wish to emphasize now is that this necessary condition is not sufficient.[239] Even if Congress would have some affirmative reason to provide an offered benefit notwithstanding the state's noncompliance with a condition, non-provision of the benefit does not amount to a proscribed penalty if the reasons that militate against providing the benefit, and that

237. For a particularly strong expression of this objection, see Bagenstos, *supra* note 136, at 35–38.

238. I read Justice Ginsburg's observation that "[c]ourts owe a large measure of respect to Congress' characterization of the grant programs it establishes" as in essence a plea for a deferential decision rule. Nat'l Fed'n of Indep. Bus. v. Sebelius, 132 S. Ct. 2566, 2636 (Ginsburg, J., dissenting in part).

239. *See supra* note 215, where the point is implied but not highlighted.

Congress treats as overriding, do not depend upon making the state's exercise of its rights more costly. And for purposes of evaluating social welfare programs, the most notable reasons that might fit this bill arise from government's legitimate interests in not exacerbating morally meaningful inequalities, and in not being party to what it takes to be morally problematic behavior, even if constitutional.

Both interests can be illustrated with a single hypothetical. Suppose that Congress offers states federal matching funds for the purpose of combatting four big killers: $W for lung cancer, $X for breast cancer, $Y for heart disease, and $Z for HIV/AIDS. State S agrees to accept the first three matching offers but not the fourth. Naturally, Congress would not be expected to provide State S with $Z for HIV/AIDS prevention, and its failure to provide that benefit would not amount to a penalty. But I'd go further. I think it plausible that Congress could refuse to provide any of the offered funds on S's refusal to provide matching funds for HIV/AIDS *even though the national interest in combatting cancer and heart disease in State S is served equally well regardless of whether that state agrees to partner with Congress to combat HIV/AIDS.* Congress might reason that State S's choices amount to morally wrongful discrimination of a sort with which it wishes not to be complicit. If such reasoning is fairly attributable to Congress, then the overriding reason for which it acts in withholding the benefit need not involve punishing or discouraging State S's exercise of its right not to participate in a federal-state program to combat HIV/AIDS. In this case, non-provision of funds for the other diseases would not run afoul of the anti-penalty principle. Possibly, on reasoning much like this, many bundled offers that *are* fairly disaggregable do not threaten to penalize rights. (Possibly, this reasoning might even save the Medicaid expansion, though my instinct is to evaluate claims of this sort with a skeptical eye lest the anti-coercion principle be too easily evaded.)

Objection 6: "What you call 'threatening a penalty,' I call 'bargaining.' It is a ubiquitous feature of commercial negotiation that, in an effort to secure a greater portion of the benefits of exchange, parties threaten not to consummate a deal on terms that they recognize would in fact serve their interests. Consider, for example, the brief story, presented earlier, of the University of Texas Law School and faculty candidate Lucy Taylor.[240] It might be that, taking opportunity costs into account, UT would genuinely prefer not to employ Taylor if she refuses to teach tax. But it might be otherwise: the school might prefer to hire her no matter what she teaches to not hiring her at all, while preferring to hire her as a tax instructor most of all. Similarly, it might prefer to hire her at an annual salary of $X to not hiring her at all, while most preferring to hire her at a salary of $X-n. On your analysis, the state actor threatens to penalize Taylor's constitutional

240. *See supra* note 171 and accompanying text.

right not to sell her labor on any particular terms if it conditions its offer of employment on her agreement to teach tax or to accept a salary lower than $X. Yet those are implausible conclusions: surely such negotiating behavior is constitutionally unobjectionable."

Response: I agree that such negotiating behavior is constitutionally unobjectionable. The state, as employer, must be entitled to bargain by means of threatening not to consummate a deal even on terms that exceed its reservation price. This is true even though its reason to carry out its threat, in the event that its conditions are not accepted, would be to vindicate the efficacy of its threats going forward. The difficult question, I think, concerns the breadth of this concession. When neither contracting party has a claim on the full transactional surplus, bargaining should be licensed precisely because there is no good way to allocate the surplus that bypasses bargaining.

But the relationships between the state and its citizens (or other persons subject to its jurisdiction), and between governments in a federal system, are different in varied ways from the relationships between private parties who contract with each other to advance their respective self-interests. Accordingly, one possible lesson from the employment hypothetical is that states are entitled, just like private parties, to haggle over transactional surpluses when acting essentially as private parties, i.e., when they are acting (more or less) as what Dormant Commerce Clause doctrine terms "market participants,"[241] but not otherwise. When the state, acting in its sovereign capacity, offers benefits to agents that hold rights against it, the interests that undergird the rights and the nature of the state's relationship to its beneficiaries might combine to direct that the offeree *does* have a claim on the full transactional surplus. Though the details of this argument remain to be worked out, I do suspect that this is at least sometimes true. And when it is, we are left without reason to accept that the state must be permitted to bargain by means of threatening penalties.

Furthermore, even to the extent that government, when *not* acting as a market participant, ought to be constitutionally permitted to "bargain" over the terms by which it distributes benefits to rightholders, it does not follow that it should enjoy the same latitude to threaten to withhold an offered benefit as do most private contracting parties. For one thing, inequalities of bargaining power loom especially large here. One plausible conclusion would be that government may not strive to secure greater benefits of exchange by threatening a penalty on terms that compel acceptance. This is not to contradict anything argued in Part II. There I argued not that compulsion is always normatively irrelevant, but only that it does not, *by itself*, have the normative significance that seven members of the *NFIB* Court attributed to it.[242] Indeed, the contract law doctrine of coercion (see subpart

241. For a good discussion, see Dan T. Coenen, *Untangling the Market-Participant Exemption to the Dormant Commerce Clause*, 88 MICH. L. REV. 395 (1989).

242. *See supra* Part II.

II(A)) saliently exemplifies that legal consequences might sensibly follow from the *conjunction* of coercion and compulsion.

Objection 7: "Is your central thesis, then, that the Medicaid expansion was unconstitutionally coercive unless, for any of several different reasons, it wasn't? If so, shouldn't you be embarrassed to have devoted fully 30,000 words to this claim?"

Response: To address these questions in reverse order: yes, and no. With respect to my latter answer, it bears emphasis that this paper is not intended as an argument that the Medicaid expansion was unconstitutional. It is intended as an analysis of the respects in which related but distinct normative concepts or principles—what I have labeled coercion and compulsion—properly bear on the constitutionality of offers of benefits conditioned on the recipient's waiver or non-exercise of a constitutional right, with a focus on conditional spending offers issued by the federal government to the states. According to the analysis I offer, some conclusions strike me as firm if not unassailable (like that the compulsion-centered reasoning that four Justices in *NFIB* put forth unequivocally is not sound[243]) whereas others are tentative. If, as I believe, there exist principles and considerations that, when combined in the right way, are fairly described as a "solution" to the conditional offer problem, that solution will not be remotely algorithmic. The most we can hope for of a proposed solution is, as Seth Kreimer counseled a generation ago, that "it at least gets the easy cases right, explains why the hard cases are hard, and allows argument to center on the appropriate factual and legal issues."[244]

Conclusion

In *National Federation of Independent Business*, the Court held, 7–2, that the Medicaid expansion provision of the Affordable Care Act amounts to unconstitutional coercion.[245] And it amounts to coercion, so the majority reasoned, because, by threatening to withhold *all* Medicaid funds from states that would decline the offer of new funds for a new class of beneficiaries, Congress presented states with a nominal choice that was functionally "no choice"—no choice because states could not rationally entertain one of the two nominal options.[246] The new conditional offer was unconstitutionally coercive, in short, because it compelled states to accept.[247]

The *NFIB* majority was half right: the Medicaid expansion probably *was* coercive in the particular sense that it compelled acceptance. But, I have argued, the majority provides no good reason to believe that *that* sense of

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

244. Kreimer, *supra* note 148, at 1301.

245. Nat'l Fed'n of Indep. Bus. v. Sebelius, 132 S. Ct. 2566, 2575 (2012) (joint opinion).

246. *Id.* at 2574–75.

247. *Id.* at 2574.

coercion is, all by itself, constitutionally meaningful, and there are powerful reasons to doubt it. If this is right, then it might seem to follow that, contrary to the majority's conclusion, the states' challenge to the Medicaid expansion gains no traction from an "anti-coercion principle." That conclusion, however, would be premature. Perhaps different meaning could be given to "coercion," and perhaps the Medicaid expansion might transgress an anti-coercion principle understood in those different terms.

In fact, there are other senses of coercion "out there," available for deployment. Normative theorists have coalesced around one in particular. According to this favored sense of coercion (and to a first pass), a conditional proposal is coercive if it would be wrongful for the maker to do as it threatens.[248] I have argued that the anti-coercion principle against which conditional offers of benefits are properly evaluated should incorporate this understanding of coercion (call it coercion, proper) and not the one that the majority employs (call it compulsion). I have also argued that embrace of the premises that conditional offers (which are also, necessarily, conditional threats) are presumptively unconstitutional when they amount to coercion, and are not presumptively unconstitutional just because they amount to compulsion, does not—contrary to prevailing scholarly wisdom—entail that conditional offers of benefits to which offerees are not legally entitled can never be unconstitutionally coercive. Withholding benefits can impermissibly penalize right holders when done in order to make exercise of a right costly or painful. Not incidentally, all of this jibes with features of the Chief Justice's reasoning that are hard to square with a superficial reading of that opinion pursuant to which compulsion does all the normative work.

I reiterate—here beating a horse that I would hope to be well-interred by this point—that my analysis of federal conditional spending is not conditional-spending particular. It depends upon two claims of far greater generality: (1) the state should not engage in the constitutional wrong of coercion, understood as conditionally threatening what would be constitutionally wrongful to do; and (2) the state may not penalize the exercise of constitutional rights in the sense of imposing adverse consequences—relative to the consequences it would otherwise impose or allow to obtain—for the purpose of punishing or discouraging the exercise of the right.

Of course, we wish to know how these general principles apply to the Medicaid expansion. I have concluded that the threat to withhold all Medicaid funds from states that would decline the offer of new funds for a new class of beneficiaries most likely does threaten to penalize the states' constitutional right to decline that offer and thus amounts to impermissible coercion. If so, the majority reached the right bottom line, though for the wrong reasons. This conclusion, though, is not ironclad. There are several

248. *See* Gunderson, *supra* note 37, at 248 (explaining that coercion involves the threat of sanctions).

possible avenues for avoiding it consistent with acceptance of the anti-coercion and anti-penalty principles as I have glossed them. For example, perhaps it is constitutionally permissible for Congress to penalize states for exercising their constitutional "privileges" or prerogatives even while it is not permissible for any level of government to penalize individuals for exercising their constitutional rights. Or perhaps a state that would accept Medicaid funding for some classes of beneficiaries but not for others would thereby exacerbate morally meaningful inequalities such that Congress might refuse to allow a state this choice for reasons that do not constitute a penalty.

Because some readers will understandably hunger for a more decisive constitutional bottom line, I will close by recommending that consumers and producers of constitutional scholarship focus more keenly than is the fashion on general principles and concepts of normative and constitutional reasoning. The application of these general principles and concepts to concrete fact patterns will frequently depend upon contestable judgments that are irreducibly subjective (to some nontrivial degree) and with respect to which constitutional theorists may lack comparative expertise. Accordingly, scholars' insistence on trying fully to resolve difficult concrete disputes predictably contributes, as Mike Seidman and Mark Tushnet diagnosed some years ago, to "the tendentious debate that has made constitutional argument so unproductive in the modern period."[249]

Perhaps, then, we should worry a little less about case-specific holdings and a little more about the state of our normative building blocks. Put more pointedly, when a court opines, say, that some action does or does not amount to coercion or to a penalty, then our first and most fundamental task is to insist, if possible, that such judgments comport with defensible accounts of the relevant concepts, and are applied consistently across cases and lines of authority (absent good reason to the contrary). We can and should appraise the job courts do in wielding the tools at their disposal. But we provide an even greater service by refining the tools.

249. SEIDMAN & TUSHNET, *supra* note 167, at 77.

Deference Lotteries

Jud Mathews*

When should courts defer to agency interpretations of statutes, and what measure of deference should agencies receive? Administrative law recognizes two main deference doctrines—the generous Chevron *standard and the stingier* Skidmore *standard—but Supreme Court case law has not offered a bright-line rule for when each standard applies.*

Many observers have concluded that courts' deference practice is an unpredictable muddle. This Article argues that it is really a lottery, in the sense the term is used in expected utility theory. Agencies cannot predict which deference standard a court will apply or with what effect, but they have a sense for how probable the different possible outcomes are. This Article presents empirical support for the "deference lottery" hypothesis, and then conducts a simple game theory analysis to understand how judicial review bears on agency behavior in statutory interpretation under deference lottery conditions.

The Article concludes that, in fact, the deference lottery can function as a flexible tool for managing agency behavior. The lottery can curb agency opportunism by imposing a risk that agencies' interpretations of statutes will face elevated scrutiny rather than Chevron *deference. This analysis offers a new perspective on deference doctrine, and in particular on the Supreme Court's* Mead *decision, which sets out the standard for when* Chevron *applies.* Mead*'s vagueness, widely derived as a bug, may in fact be a feature. Still, the deference lottery can backfire badly if* Skidmore *is applied too stringently, as the Article shows.*

---

* Assistant Professor, Penn State University Dickinson School of Law. My thanks to Amitai Aviram, Ian Ayres, Nuno Garoupa, Kristin Hickman, David Hyman, Richard Kaplan, David Kaye, Kurt Lash, Robert Lawless, Larry Ribstein, Richard Ross, Arden Rowell, David Schraub, Jamelle Sharpe, Paul Stancil, Suja Thomas, Tom Ulen, Verity Winship, and all the participants in the Big Ten UnTenured Conference and the University of Illinois College of Law workshop for helpful comments on this project. I also thank the editors at the *Texas Law Review* for their care and attention in the editorial process.

I. Introduction

When should courts defer to agency interpretations of statutes and what measure of deference should agencies receive? Administrative law recognizes two main deference doctrines—the generous *Chevron*[1] standard and the stingier *Skidmore*[2] standard[3]—but Supreme Court case law has not offered a bright-line rule for which standard applies when.[4] Further, even when a court purports to operate *within* a given deference regime, it is not clear that the standards are applied consistently from case to case.[5] Empirical work has confirmed that courts often fail to apply deference standards in circumstances where their own doctrine indicates they should.[6] Moreover, courts continue to apply other deference doctrines in special contexts, driving

1. Chevron U.S.A. Inc. v. Natural Res. Def. Council, Inc., 467 U.S. 837 (1984).

2. Skidmore v. Swift & Co., 323 U.S. 134 (1944).

3. Under *Chevron*, courts are to accept any "permissible" (meaning reasonable) agency construction of an ambiguous statute. *Chevron*, 467 U.S. at 843. When *Skidmore* applies, a court gives deference on a sliding scale: an agency's interpretation will be credited in proportion to its "power to persuade." *Skidmore*, 323 U.S. at 140. The standards are discussed in more detail below. *See infra* Part II.

4. United States v. Mead Corp., 533 U.S. 218, 229–31 (2001) (holding that *Chevron* deference is due when it is "apparent" that "Congress would expect the agency to be able to speak with the force of law when it addresses ambiguity in the statute or fills a space in the enacted law," but declining to set out conclusive criteria for establishing the requisite congressional intent).

5. *See, e.g.*, Jack M. Beermann, *End the Failed* Chevron *Experiment Now: How* Chevron *Has Failed and Why It Can and Should Be Overruled*, 42 CONN. L. REV. 779, 809–22 (2010) (detailing inconsistencies in the application of *Chevron*); Lisa Schultz Bressman, *How* Mead *Has Muddled Judicial Review of Agency Action*, 58 VAND. L. REV. 1443, 1458–64 (2005) (describing inconsistencies in when appeals courts apply different deference doctrines).

6. William N. Eskridge, Jr. & Lauren E. Baer, *The Continuum of Deference: Supreme Court Treatment of Agency Statutory Interpretations from* Chevron *to* Hamdan, 96 GEO. L.J. 1083, 1090 (2008); Connor N. Raso & William N. Eskridge, Jr., Chevron *as a Canon, Not a Precedent: An Empirical Study of What Motivates Justices in Agency Deference Cases*, 110 COLUM. L. REV. 1727, 1740 (2010).

the predictability of judicial practice further down.[7] Taken together, all this means that agencies seeking to defend statutory interpretations in court can anticipate with confidence neither what standard will be applied nor how the court will apply it.

The confused state of deference doctrine has attracted its share of critical commentary.[8] The Supreme Court's 2001 *United States v. Mead Corp.*[9] decision, which declined to mark off the border between *Chevron*'s domain and *Skidmore*'s with a bright-line rule, has been a focal point for criticism.[10] To be sure, a lack of clarity over the scope of deference an agency interpretation will receive—an unpredictability in the law generally—imposes costs.[11] Here, the costs of an unpredictable deference

---

7. Although *Chevron* and *Skidmore* are the deference standards most often employed, the Court has articulated a number of other deference standards for use in specialized contexts. *See* Eskridge & Baer, *supra* note 6, at 1090 (identifying five distinct modes of deference to agency interpretations, including *Seminole Rock*, *Curtiss-Wright*, and *Beth Israel*). Work on deference doctrines in the lower federal courts has revealed a similarly variegated picture. *See* Jason J. Czarnezki, *An Empirical Investigation of Judicial Decisionmaking, Statutory Interpretation, and the* Chevron *Doctrine in Environmental Law*, 79 U. COLO. L. REV. 767, 770–71 (2008) (observing that "there remains much confusion and conflation in the circuits over how to apply the *Chevron* doctrine").

8. *See, e.g.*, ADRIAN VERMEULE, JUDGING UNDER UNCERTAINTY: AN INSTITUTIONAL THEORY OF LEGAL INTERPRETATION 215–16 (2006) (characterizing *Mead* as "close to disastrous on institutional grounds" owing to the "cognitive and institutional load that the increasing complexity of *Mead*'s legal regime imposes on lower courts, litigants, and other actors"); David J. Barron & Elena Kagan, Chevron*'s Nondelegation Doctrine*, 2001 SUP. CT. REV. 201, 205 (arguing that "the Court's reliance on congressional intent should give way to a frankly policy-laden assessment of the appropriate allocation of power in the administrative state" and "that the underlying policy evaluation of the Court misidentifies the criteria that should govern this allocation by focusing on the presence of formal procedures and generality"); Beermann, *supra* note 5, at 788–835 (detailing inconsistencies in the application of *Chevron*); Lisa Schultz Bressman, Chevron*'s Mistake*, 58 DUKE L.J. 549, 549 (2009) (contending that *Chevron* "asks courts to determine whether Congress has delegated to administrative agencies the authority to resolve questions about the meaning of statutes that those agencies implement, but . . . does not give courts the tools for providing a proper answer"); William S. Jordan, III, *Judicial Review of Informal Statutory Interpretations: The Answer is* Chevron *Step Two, Not* Christensen *or* Mead, 54 ADMIN. L. REV. 719, 719 (2002) (describing the Court's current approach to the review of administrative agencies' informal statutory interpretations as "a cumbersome, unworkable regime under which courts must draw increasingly fine distinctions using impossibly vague standards").

9. 533 U.S. 218 (2001).

10. *See, e.g.*, Bressman, *supra* note 5, at 1143–44 (endorsing the view that *Mead* caused judicial review of agency action to "devolve into chaos"); William S. Jordan, III, United States v. Mead: *Complicating the Delegation Dance*, 31 ENVTL. L. REP. 11425, 11425 (2001) (opining that *Mead* obscured *Chevron*'s "treasured clarity"); Thomas W. Merrill, *The* Mead *Doctrine: Rules and Standards, Meta-Rules and Meta-Standards*, 54 ADMIN. L. REV. 807, 809 (2002) (arguing that both the majority and the dissent in *Mead* were mistaken); Adrian Vermeule, *Introduction:* Mead *in the Trenches*, 71 GEO. WASH. L. REV. 347, 347 (2003) (arguing that the flaws and incoherencies in the case law applying *Mead* "are traceable to the flaws, fallacies, and confusions of the *Mead* decision itself").

11. Foundational works on the effects of uncertain legal standards include Richard Craswell & John E. Calfee, *Deterrence and Uncertain Legal Standards*, 2 J.L. ECON. & ORG. 279 (1986) and Gillian K. Hadfield, *Weighing the Value of Vagueness: An Economic Perspective on Precision in*

regime might include increased litigation,[12] more agency reversals in court,[13] "defensive rulemaking" on the part of agencies,[14] or perhaps a move away from rulemaking entirely.[15] A fuller accounting of our deference practice, however, should consider whether unpredictability might yield benefits as well as costs. This Article begins that work.

The key to this Article's unique contributions is the insight that agencies face a "deference lottery" when they advance a statutory interpretation in a notice-and-comment rulemaking or formal adjudication.[16] The Article uses the term "lottery" in the sense it is used in expected utility theory. A person faces a lottery any time he or she does not know what the outcome of a process will be, but does know what the different possible outcomes are and what the probability of each is.[17] In more formal terms, a lottery refers to any discrete probability distribution over outcomes.[18] For instance, if I buy a scratch-off lottery ticket, obviously I do not know what its payoff will be, but I do know the odds (if I read the fine print on the back). For instance, the ticket may pay $1,000 with a probability of 1-in-10,000, $1,000,000 with a probability of 1-in-10,000,000, and $0 with a probability of 9,989,999-in-10,000,000.[19] We frequently encounter lotteries outside of the gaming context as well. For instance, based on historical averages, we expect that an A-rated municipal bond will pay its face value with .97 probability and default with .03 probability.[20]

---

the Law, 82 CALIF. L. REV. 541 (1994). Notable recent works include Yuval Feldman & Shahar Lifshitz, *Behind the Veil of Legal Uncertainty*, LAW & CONTEMP. PROBS., Spring 2011, at 133.

12. *See Mead*, 533 U.S. at 250 (Scalia, J., dissenting) ("[I]n an era when federal statutory law administered by federal agencies is pervasive, and when the ambiguities (intended or unintended) that those statutes contain are innumerable, totality-of-the-circumstances *Skidmore* deference is a recipe for uncertainty, unpredictability, and endless litigation.").

13. *Id.*

14. *See* Jerry L. Mashaw, *Improving the Environment of Agency Rulemaking: An Essay on Management, Games, and Accountability*, LAW & CONTEMP. PROBS., Spring 1994, at 185, 203 (referring not to the deference lottery, but to the unpredictability generally engendered by aggressive judicial review of agency rulemakings).

15. *Id.*

16. These are situations where, after *Mead*, *Chevron* applies presumptively but not definitively. As a shorthand, I sometimes refer to "agency statutory interpretations" to mean statutory interpretations rendered in these formats. The Article focuses on this subset of agency statutory interpretations because the most important agency decisions are likely to be taken pursuant to one of these procedures, as opposed to less formal forms of agency action.

17. This somewhat technical usage is uncommon, though not unknown, in legal scholarship. *See generally* Adrian Vermeule, *The Delegation Lottery*, 119 HARV. L. REV. F. 105 (2006).

18. *See* MARTIN J. OSBORNE, AN INTRODUCTION TO GAME THEORY 501 (2004) ("We refer to a probability distribution over outcomes as a *lottery* over outcomes."). For a more in-depth and technical discussion, see NOLAN MCCARTY & ADAM MEIROWITZ, POLITICAL GAME THEORY: AN INTRODUCTION 27–33 (2007).

19. For the actual odds from a popular multistate lottery, see *Powerball–Prizes and Odds*, POWERBALL, http://www.powerball.com/powerball/pb_prizes.asp.

20. H.R. REP. NO. 110-835, at 5 (2008).

How does the lottery concept translate to the administrative law context? I argue that courts' deference practice contains *two distinct sources* of unpredictability that combine to generate a lottery with distinctive features. When an agency advances a statutory interpretation in a notice-and-comment rulemaking or formal adjudication, it can reliably predict neither which standard of review will be applied to its statutory interpretation—*Chevron* or *Skidmore*—nor, if *Chevron* deference is not granted, whether or not its interpretation will survive review. However, from observing judicial practice and doctrine, the agency can have a fairly good sense for both the *ex ante* odds of getting *Chevron* review and for the odds a given interpretation would survive *Skidmore* review.

In the terminology of this Article, agencies thus face a deference lottery that is a composite of *two* lotteries, which I refer to as the *Chevron* lottery and the *Skidmore* lottery.[21] The core idea is that an agency interpretation faces some probability of receiving *Chevron* deference on review (the *Chevron* lottery), and in the event that *Chevron* is not forthcoming, some probability of surviving judicial scrutiny under *Skidmore* (the *Skidmore* lottery). This structure gives courts two levers over agencies: they can tweak the *Chevron* lottery—altering the probability that agencies will be reviewed under *Chevron*—and the *Skidmore* lottery—adjusting the stringency of review within the *Skidmore* framework when *Chevron* analysis is not forthcoming.

Of course, it is not only deference law that could be characterized as a lottery. Laws are rarely fully determinate and the outcomes of judicial processes can almost never be predicted with certainty. But the lottery characterization is especially apt here, as the quantum of unpredictability in deference issues is especially high.[22] Moreover, the deference lottery has some distinctive and interesting properties, owing to the structure of deference doctrine.

This Article explores what follows if we take seriously the idea that, from the perspective of agencies, the deference regime is a lottery. It makes three significant contributions. First, it provides empirical evidence that the "deference lottery" is a reasonable characterization of how agencies experience judicial review. Part II surveys the development of deference doctrine, highlighting its sources of uncertainty, and then examines how deference is actually practiced in the courts. Drawing on existing empirical studies of deference practice in both the Supreme Court and the federal appellate courts, this Article identifies evidence that a deference lottery with the features described here approximates the environment that agencies

21. The term for a lottery with outcomes that are themselves lotteries is "compound lottery." CHRISTIAN GOLLIER, THE ECONOMICS OF RISK AND TIME 4 (2001).

22. *See* Eskridge & Baer, *supra* note 6, at 1091 (observing that "there is no clear guide as to when the Court will invoke particular deference regimes, and why").

actually face on judicial review.[23] This is the first work to characterize courts' deference practice in these terms and to offer support for the claim.

Second, the Article uses the concept of the deference lottery to unlock new insights into how unpredictability in our deference regime can reward or punish agency behavior in important, and sometimes counterintuitive, ways. The method this Article adopts is to explore the dynamics of the deference lottery using a simple model of agency–court interactions. This approach adopts the perspective of Principal–Agent (PA) theory.[24] It treats the agency as the agent of an enacting Congress, tasked with carrying out a statutory regime but subject to the classic agency problem: the agency's preferences may diverge from those encoded in the statute and the enacting Congress—the principal—has limited tools for keeping the agency on track.[25] On this view, judicial review of agency statutory interpretations is understood as a strategy for monitoring agency performance.

The stylized model of the deference lottery developed in this Article generates surprising implications for administrative law. The first is that, relative to a *Chevron*-only regime, the deference lottery offers a more flexible tool for shaping agency behavior. A deference lottery can encourage a rational agency to choose an interpretation that lies somewhere between the safest and the most adventurous version that the agency can hope to get away with. This Article takes no position on what kind of interpretation is best—that is, on what is the optimal level of agency slack in the statutory interpretation context. Rather, the Article shows that a deference lottery opens possibilities for shaping agency behavior that are not available under a *Chevron*-only regime. In this way, the Article casts a new, and more favorable, light on *Mead*. To the extent that unpredictability in the deference regime can have desirable effects, the much-maligned vagueness of *Mead* may be a feature, not a bug.

The Article also shows how subtle changes to the deference lottery could have pronounced—and undesirable—effects on agency behavior. One of the most striking findings is that, paradoxically, increasing the scrutiny an agency will receive under *Skidmore* can actually encourage an agency to adopt a *less* faithful interpretation of the statute. The reason is this: if *Skidmore* deference is very difficult to satisfy, at a certain point, the expected rewards from compromising on policy are so meager that it makes sense for an agency to give up its effort to "win" the *Skidmore* lottery entirely. Instead, the expected benefit is higher from selecting an interpretation the

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

24. For the classic introduction to Principal–Agent theory, as relevant to the public law context, see Terry M. Moe, *The New Economics of Organization*, 28 AM. J. POL. SCI. 739, 756–58 (1984).

25. For further discussion of the PA logic at work in this Article, and of how the analysis accommodates the President's role in executing statutes, *see infra* Part III.

agency prefers and "betting it all" on the *Chevron* lottery.[26] A strategy to avoid this outcome is to construct a deference regime in which a fairly deferential *Skidmore* standard is applied fairly frequently. The Article offers some reasons to think that this is the kind of deference regime federal agencies face.

The third major contribution of the Article is to the body of work on uncertainty and risk in the law generally.[27] This piece builds on a research agenda that has identified, in general terms, both the potential behavior-shaping value of vagueness in the law[28] and some of its limits.[29] This Article advances the state of scholarship with respect to both the scope of application and the development of theory. This is the first piece to explore at length how the unpredictability of the deference regime in administrative law bears on agencies' strategies of statutory interpretation. As such, it brings a new perspective to the extensive legal literature on judicial review of agency statutory interpretation.

Moreover, the Article explores the dynamics of a doctrinal structure that generalizes beyond administrative law. The scenario this Article analyzes is one where a court will evaluate conduct within one of two possible regimes, where one is relatively permissive and the other relatively stringent, and the decision which regime applies is governed by a vague standard (here, the *Mead* test). This research in principle translates to other doctrinal settings with the same features: for instance, to corporate law, where courts possess two standards that could plausibly be used to evaluate certain actions of directors and officers—the forgiving business judgment rule and the strict duty of loyalty[30]—and the standard for when each applies is opaque.[31]

---

26. The point has some parallels to Richard Craswell and John E. Calfee's conclusion about the deterrent value of unpredictability. *See* Craswell & Calfee, *supra* note 11, at 287 (positing a counterintuitive, inverse relation between certainty and compliance incentives in certain criminal law contexts).

27. Economists often make a distinction between risk and uncertainty. FRANK H. KNIGHT, RISK, UNCERTAINTY AND PROFIT 197–232 (Univ. of Chi. Press 1971) (1921). Actors face risk when they do not know which event will occur, but they know the relative probabilities of the possible events; they face uncertainty when they do not even know the probabilities. *See* Daniel A. Farber, *Uncertainty*, 99 GEO. L.J. 901, 903 (2011) (reiterating this distinction); Sarah B. Lawsky, *Probably? Understanding Tax Law's Uncertainty*, 157 U. PA. L. REV. 1017, 1026–31 (2009) (building on this dichotomy). This distinction is not always drawn in the law and economics scholarship, however, and for simplicity of exposition and consistency with ordinary usage, this Article sometimes refers to the "uncertainty" in deference doctrine even though its analysis supposes that the frequencies of different outcomes are knowable.

28. *See* Hadfield, *supra* note 11, at 548–49 (suggesting that vague legal standards might elicit a more socially desirable mix of behaviors than determinate standards).

29. *See* Craswell & Calfee, *supra* note 11, at 287 (finding that, contrary to the authors' prior conjecture, "it is not necessarily true that reductions in the level of uncertainty will improve [defendants'] compliance decisions").

30. On the development of these standards, see Marcia M. McMurray, Note, *An Historical Perspective on the Duty of Care, the Duty of Loyalty, and the Business Judgment Rule*, 40 VAND. L. REV. 605, 606–18, 623–28 (1987).

The rest of the Article is structured as follows. Part II combines doctrinal and empirical analysis to show that the deference lottery concept reasonably approximates how agencies experience judicial review of their statutory interpretations. Part III develops a simple model of how the configuration of the deference lottery can shape agencies' strategies of statutory interpretation, and presents the key results. Part IV considers the applicability and limitations of the Part III analysis, suggests some implications for administrative law, and concludes.

II. Deference in Doctrine; Deference in Fact

This Part establishes that the judicial review environment for agency statutory interpretations can meaningfully be characterized as a deference lottery. In other words, in any individual case, agencies cannot predict with confidence either what standard the court will apply or, in the event sliding-scale *Skidmore* deference is applied, whether or not the court will uphold its interpretation, but the relative frequencies of the different possible outcomes are fairly stable over time. This Article does not claim that agencies face a perfect lottery, where the odds are known with certainty, and certainly does not claim that courts make their deference decisions by choosing at random from among the possible outcomes.[32] Below, I detail the ways in which the experience of agencies facing judicial review may diverge from a true lottery.[33] What I am claiming is that, *from the perspective of the agency*, the experience of judicial review in the statutory interpretation context reasonably approximates a lottery.

A. Deference Doctrine

This subpart summarizes the development of deference doctrine in administrative law. This story has been told before, sometimes in great detail.[34] This account emphasizes a persistent theme: the Supreme Court's oscillation between clarity and obscurity in deference standards. More than once, the Supreme Court has established a fairly straightforward policy regarding the deference due to agencies, only to chafe under its rigidity and introduce more nuance.

The early decades of modern administrative law saw the Supreme Court swing from one pole to the other in its approach towards statutory

---

31. *Cf.* Melvin Aron Eisenberg, *The Divergence of Standards of Conduct and Standards of Review in Corporate Law*, 62 FORDHAM L. REV. 437, 461–67 (1993) (clarifying the various standards applied by courts to corporate decisions and elaborating a framework for understanding the methodology by which judges apply these standards). Thanks to Larry Ribstein for this point.

32. The Article, however, considers reasons why it is not surprising that judicial behavior in the aggregate approximates a lottery. *See infra* subpart II(C).

33. *See infra* section II(B)(1).

34. *See, e.g.*, Reuel E. Schiller, *The Era of Deference: Courts, Expertise, and the Emergence of New Deal Administrative Law*, 106 MICH. L. REV. 399 (2007) (recounting the development of deference in administrative law since the New Deal).

interpretation by agencies. The Court reserved for itself the authority to define key terms of regulatory statutes, such as "unfair methods of competition in commerce" (in the FTC Act), into the 1920s,[35] but by the New Deal era had largely adopted a policy of broad deference to agency interpretations.[36] In *Gray v. Powell*,[37] the Supreme Court heard a challenge to the National Bituminous Coal Commission's determination that the Seaboard Air Line Railway Company was a consumer of coal only, and not also a "producer" within the meaning of the Bituminous Coal Code.[38] The Court sharply limited the scope of its own review of the agency's interpretation:

> In a matter left specifically by Congress to the determination of an administrative body, . . . the function of review placed upon the courts . . . is fully performed when they determine that there has been a fair hearing, with notice and an opportunity to present the circumstances and arguments to the decisive body, and an application of the statute in a just and reasoned manner.
>
> . . . .
>
> Where, as here, a determination has been left to an administrative body, this delegation will be respected and the administrative conclusion left untouched. . . .
>
> . . . Just as in the *Adkins* case the determination of the sweep of the term "bituminous coal" was for this same administrative agency, so here there must be left to it, subject to the basic prerequisites of lawful adjudication, the determination of "producer."[39]

Gray seems to demote the Court, leaving it only to check that the agency's interpretation followed proper process and represented a "just and reasoned" application of the statute.

The Court took a similar approach in its well-known decision three years later in *NLRB v. Hearst Publications, Inc.*,[40] upholding the NLRB's determination that newsboys are "employees" within the meaning of the National Labor Relations Act.[41] While reserving to the courts a leading role

35. FTC v. Gratz, 253 U.S. 421, 427 (1920).

36. For a detailed account, see Schiller, *supra* note 34, at 407–12, 430–38. As Schiller notes, the Court's practice was not entirely uniform in any period; for instance, the Court more vigorously policed the activity of agencies operating on the periphery of traditional police powers. *Id.* at 407–09; *see also* Reuel Schiller, *"Saint George and the Dragon": Courts and the Development of the Administrative State in Twentieth-Century America*, 17 J. POL'Y HIST. 110, 113 (2005) (arguing that courts were most deferential to agencies in "areas of regulation that fit comfortably within a traditional reading of the police powers").

37. 314 U.S. 402 (1941).

38. *Id.* at 403–06.

39. *Id.* at 411–13 (citations omitted).

40. 322 U.S. 111 (1944).

41. *Id.* at 132.

in questions of abstract statutory interpretation,[42] "where the question is one of specific application of a broad statutory term in a proceeding in which the agency administering the statute must determine it initially, the reviewing court's function is limited."[43] More specifically, "the Board's determination that specified persons are 'employees' under this Act is to be accepted if it has 'warrant in the record' and a reasonable basis in law."[44] Even more striking, as Reuel Schiller has noted, federal appeals courts during the same period frequently deferred to agency interpretations on matters of constitutional law, most notably, First Amendment issues raised by administrative practices.[45] In the immediate aftermath of the New Deal, federal courts thus generally adopted a policy of broad deference to agency statutory interpretations.[46]

The Court's decision in *Skidmore v. Swift & Co.*,[47] decided later in 1944, reflects a more contextual approach to deference. Here, the question was whether nights that private firefighters spent on call and on premises at the Swift plant counted as "working time" for purposes of the Fair Labor Standards Act.[48] The statute routed disputes under the Act to the courts, not the Labor Department, but the Administrator of the Wage and Hour Division had issued an "Interpretive Bulletin" containing a standard for calculating working time and filed an amicus curiae brief on the firemen's behalf.[49] The district court had reviewed the question de novo, not taking the agency's position into account.[50] In an opinion by Justice Jackson, the Supreme Court ruled that the district court had failed to give proper consideration to the agency's views on the subject, and remanded.[51] How much deference was owed exactly? The Court offered this formula:

---

42. *See id.* at 130–31 ("Undoubtedly questions of statutory interpretation, especially when arising in the first instance in judicial proceedings, are for the courts to resolve, giving appropriate weight to the judgment of those whose special duty it is to administer the questioned statute.").

43. *Id.* at 131. For a similar, contemporaneous approach with language that prefigures *Chevron*, see *Dobson v. Commissioner of Internal Revenue*, 320 U.S. 489 (1943). There the Court held that the tax court's interpretation of whether certain settlement proceeds qualified as "income" need only "have 'warrant in the record' and a rational basis in the law." *Id.* at 501.

44. *Hearst*, 322 U.S. at 131.

45. *See* Schiller, *supra* note 34, at 436–38 (noting circuit court deference to NLRB decisions punishing employers for statements made during union elections); Reuel E. Schiller, *Free Speech and Expertise: Administrative Censorship and the Birth of the Modern First Amendment*, 86 VA. L. REV. 1, 96–101 (2000) (arguing that scarcity of available broadcast stations and a belief in agency expertise led the Supreme Court to defer to the FCC's content-based regulation of speech).

46. *See* Schiller, *supra* note 34, at 429–38 (discussing judicial deference to administrative findings in the 1930s and 1940s).

47. 323 U.S. 134 (1944).

48. *Id.* at 135–36.

49. *Id.* at 138–39.

50. *See id.* at 140 ("[A]lthough the District Court referred to the Administrator's Bulletin, its evaluation and inquiry were apparently restricted by its notion that waiting time may not be work, an understanding of the law which we hold to be erroneous.").

51. *Id.*

> We consider that the rulings, interpretations and opinions of the Administrator under this Act, while not controlling upon the courts by reason of their authority, do constitute a body of experience and informed judgment to which courts and litigants may properly resort for guidance. The weight of such a judgment in a particular case will depend upon the thoroughness evident in its consideration, the validity of its reasoning, its consistency with earlier and later pronouncements, and all those factors which give it power to persuade, if lacking power to control.[52]

The *Skidmore* formula tailors the deference owed to the infinite variety of individual cases. From early on, *Skidmore* has divided opinion between those who appreciate its sensitivity to context[53] and those who lament its open-endedness and question how it is to be administered consistently.[54]

The Supreme Court continued to refine its deference jurisprudence,[55] but "revolutionary" change[56] came forty years after *Skidmore*, in *Chevron U.S.A. Inc. v. National Resources Defense Council, Inc.*[57] In deciding whether to permit the EPA's interpretation of "stationary source" to apply to entire facilities rather than individual smokestacks,[58] the Court inaugurated its famous two-step approach to deference decisions:

> When a court reviews an agency's construction of the statute which it administers, it is confronted with two questions. First, always, is the question whether Congress has directly spoken to the precise question at issue. If the intent of Congress is clear, that is the end of the matter; for the court, as well as the agency, must give effect to the unambiguously expressed intent of Congress. If, however, the court

---

52. *Id.*

53. *See* Reginald Parker, *Administrative Interpretations*, 5 MIAMI L.Q. 533, 538 (1951) (describing *Skidmore* as "the golden middle" between approaches that abdicate courts' authority to review or usurp agencies' authority to interpret.).

54. *See, e.g.*, United States v. Mead Corp., 533 U.S. 218, 250 (2001) (Scalia, J., dissenting) ("*Skidmore* deference is a recipe for uncertainty, unpredictability, and endless litigation."); Melissa Hart, *Skepticism and Expertise: The Supreme Court and the EEOC*, 74 FORDHAM L. REV. 1937, 1945 (2006) (noting the "open-ended and malleable" nature of the *Skidmore* standard).

55. For examples of the Court's less-structured post-*Skidmore*, pre-*Chevron* deference cases, see Beth Israel Hosp. v. NLRB, 437 U.S. 483, 500–01 (1978), Nat'l Muffler Dealers Ass'n v. United States, 440 U.S. 472, 476–77 (1979), and Batterton v. Francis, 432 U.S. 416, 425 (1977).

56. Thomas W. Merrill and Kristin E. Hickman have noted, but do not themselves subscribe to, the common view in administrative law scholarship that *Chevron* amounted to a revolution. *See* Thomas W. Merrill & Kristin E. Hickman, Chevron*'s Domain*, 89 GEO. L.J. 833, 834–35 (2001) ("[W]e accept for present purposes the *Chevron* revolution as an established fact . . . ."). They and others have emphasized *Chevron*'s roots in existing precedents. *See id.* at 833 ("The idea that deference on questions of law is sometimes required was not new."). Indeed, already in the late 1940s, Professor Nathaniel Nathanson had identified a nascent doctrine "which teaches that there are occasions when the reviewing court need not be persuaded that the administrative agency's choice of conflicting interpretations is right, but only that it is reasonable." Nathaniel L. Nathanson, *Administrative Discretion in the Interpretation of Statutes*, 3 VAND. L. REV. 470, 470 (1950).

57. 467 U.S. 837 (1984).

58. *Id.* at 840.

> determines Congress has not directly addressed the precise question at issue, the court does not simply impose its own construction on the statute, as would be necessary in the absence of an administrative interpretation. Rather, if the statute is silent or ambiguous with respect to the specific issue, the question for the court is whether the agency's answer is based on a permissible construction of the statute.[59]

On the face of it, *Chevron* appeared to jettison the complexities of *Skidmore* in favor of two yes-or-no questions: is the statute ambiguous, and if so, is the agency's interpretation "permissible," i.e., reasonable?[60] *Chevron* attracted limited notice at first,[61] but after its enthusiastic adoption by the D.C. Circuit[62] and its increasing popularity on the Supreme Court itself,[63] its potential to broaden the scope of deference and simplify the analysis quickly became apparent. But *Chevron*'s simple formula concealed difficult questions,[64] including the matter of "Step Zero": the question of *Chevron*'s scope.[65] Does *Chevron* apply to every statutory interpretation advanced by an agency or only some subset? And what analysis governs cases that do not get *Chevron* review?

As the Court took up these questions, the contours of deference doctrine became harder to follow. In *Christensen v. Harris County*,[66] the Court voiced a hard-line approach to *Chevron*'s scope: "Interpretations such as those in opinion letters—like interpretations contained in policy statements, agency manuals, and enforcement guidelines, all of which lack the force of law—do not warrant *Chevron*-style deference."[67] Rather, they are analyzed under *Skidmore*.[68]

The following term, in *Mead*, the Supreme Court rejected *Christensen*'s proposition that the test for *Chevron* was the formality of the agency pronouncement. Returning to the stated rationale for deference in *Chevron*, Justice Souter in *Mead* wrote that *Chevron* deference is due whenever

---

59. *Id.* at 842–43 (footnotes omitted).

60. *See id.* at 844 ("In such a case, a court may not substitute its own construction of a statutory provision for a reasonable interpretation made by the administrator of an agency.").

61. For his part, Justice Stevens, the author of *Chevron*, regarded the opinion as merely a restatement of existing law. *See* Thomas W. Merrill, *The Story of* Chevron*: The Making of an Accidental Landmark*, *in* ADMINISTRATIVE LAW STORIES 399, 420 & n.76 (Peter L. Strauss ed., 2006).

62. *Id.* at 422–23.

63. *Id.* at 421–23.

64. *See* Merrill & Hickman, *supra* note 56, at 848–52 (cataloguing questions about *Chevron*'s scope). Even in cases where it is clear that *Chevron* applies, the core terms of the *Chevron* test are hardly self-defining: what does it mean for a statute to be "ambiguous" or for an interpretation to be "reasonable"?

65. Cass R. Sunstein, Chevron *Step Zero*, 92 VA. L. REV. 187, 191 (2006).

66. 529 U.S. 576 (2000).

67. *Id.* at 587.

68. *Id.*

"Congress would expect the agency to be able to speak with the force of law when it addresses ambiguity in the statute or fills a space in the enacted law."[69] The formality of the agency's interpretation is relevant to this inquiry, but not dispositive. Authorization from Congress to engage in rulemaking or adjudication is "a very good indicator of delegation meriting *Chevron* treatment,"[70] but "we have sometimes found reasons for *Chevron* deference even when no such administrative formality was required and none was afforded."[71] The key question, though, is always one of congressional intent: did Congress mean for the agency, rather than the court, to be interpreting the statute? *Mead* offered little guidance, however, into how that inquiry should be conducted. Justice Souter catalogued indicia of classification rulings' informality before pronouncing them "beyond the *Chevron* pale,"[72] but did not explain which features, if any, were dispositive. As Thomas Merrill concludes, *Mead* offers "an undefined standard that invites consideration of a number of variables of indefinite weight."[73] *Mead* also confirmed that when *Chevron* does not apply, *Skidmore* deference does.[74]

Mead's majority opinion prompted a scathing dissent from Justice Scalia, who predicted that its "wonderfully imprecise" test would generate "protracted confusion."[75] *Mead*'s reception among administrative law scholars was scarcely more hospitable.[76] The opinion was judged "opaque even by Justice Souter's standards"[77] and faulted for "provid[ing] little guidance to lower courts, agencies, and regulated parties about how to discern congressional intent in any given set of circumstances."[78] The Court's 2002 decision in *Barnhart v. Walton*[79] did little to clear up the confusion. That case concerned an interpretation of the Social Security Act that the Social Security Administration advanced first in a variety of informal formats over several decades and ultimately in a notice-and-comment regulation.[80] In determining that *Chevron* provided the correct standard of

69. United States v. Mead Corp., 533 U.S. 218, 229 (2001).

70. *Id.*

71. *Id.* at 231.

72. *Id.* at 234. The features Justice Souter noted included the large number of rulings issued each year from multiple offices, the rulings' limited precedential value, and the lack of supporting evidence in the legislative history that these rulings were intended to have binding legal authority. *Id.* at 231–34.

73. Merrill, *supra* note 10, at 813.

74. *Mead*, 533 U.S. at 234–38.

75. *Id.* at 245 (Scalia, J., dissenting).

76. *See supra* note 10 and accompanying text.

77. Vermeule, *supra* note 10, at 347.

78. Thomas W. Merrill & Kathryn Tongue Watts, *Agency Rules with the Force of Law: The Original Convention*, 116 HARV. L. REV. 467, 480 (2002).

79. 535 U.S. 212 (2002).

80. *Id.* at 217, 219–20.

deference, Justice Breyer's decision for the Court considered a laundry list of factors:

> In this case, the interstitial nature of the legal question, the related expertise of the Agency, the importance of the question to administration of the statute, the complexity of that administration, and the careful consideration the Agency has given the question over a long period of time all indicate that *Chevron* provides the appropriate legal lens through which to view the legality of the Agency interpretation here at issue.[81]

Barnhart thus introduced a set of factors to govern the threshold question of whether *Chevron* deference applies that resembles, at least loosely, the *Skidmore* analysis.[82] Though *Barnhart* has been cited hundreds of times by the lower courts, in none of the eight cases in which the Supreme Court cites to *Barnhart*[83] does the Court in fact follow its full framework for *Chevron* Step Zero analysis.

Subsequent cases have made further refinements to the deference regime,[84] but have not changed its basic architecture. Whether an agency statutory interpretation warrants *Chevron* or *Skidmore* deference, then, turns on an inquiry into congressional intent: did Congress mean for the agency to be able to decide the issue at hand with the force of law? If so, *Chevron* governs, and if not, *Skidmore* does. The Court has neither repudiated nor consistently applied its framework from *Barnhart*, which incorporates agency expertise, the scope of the legal question, the length of the agency's experience, and other factors into this threshold determination.[85] In the event that *Skidmore* applies, the agency stands a better shot of surviving review the more persuasive its interpretation is.

B. The Data on Courts' Deference Practices

This subpart turns from doctrine to data, to consider what is known about how the *Chevron*/*Skidmore* regime functions in practice. It draws from several important empirical studies of judicial review of agency statutory interpretations that have appeared in the past five years. The data and analysis assembled in this subpart are not comprehensive, and cannot answer all questions about how the deference regime operates in practice. But they offer a picture, however incomplete, of the deference landscape that agencies

---

81. *Id.* at 222.

82. *See supra* text accompanying note 52.

83. As of April 2, 2013.

84. Many commentators had concluded, in the wake of *Mead*, that agency use of formal adjudication or notice-and-comment rulemaking procedures were sufficient, if not necessary, to guarantee *Chevron* deference. *See, e.g.*, Sunstein, *supra* note 65, at 218. In his *Brand X* concurrence, Justice Breyer clarified his view that such formal procedures were not sufficient to guarantee *Chevron* deference. *See* Nat'l Cable & Telecomms. Ass'n v. Brand X Internet Servs., 545 U.S. 967, 1004 (2005) (Breyer, J., concurring).

85. *See supra* text accompanying note 83.

face. And what they show is broadly consistent with this Article's claim that the courts' practice amounts to a deference lottery.

1. Deference in the Supreme Court.—Not surprisingly, deference practice in the Supreme Court has been the object of the most intensive study, starting with the landmark 1990 article by Peter Schuck and Donald Elliott.[86] The most comprehensive source on Supreme Court deference to agency statutory interpretations yet assembled is a dataset constructed by William Eskridge and Lauren Baer.[87] Eskridge and Baer studied every Supreme Court case involving an agency interpretation of a statute between the *Chevron* decision and the end of the 2005 Term, 1,014 in all.[88] They coded each case on 156 variables that capture most of the issues one would expect to be relevant to the deference given by the Court.[89] This dataset is the richest available for exploring how deference plays out in practice, and it forms the basis for the analysis here.

That said, the fact that the data is from Supreme Court cases only limits the generalizability of the results. First, while Supreme Court opinions of course are the most authoritative, as a matter of volume the Supreme Court has a much less active role in shaping the contours of administrative law than the lower courts.[90] To the extent that agencies shape their behavior in response to cues they get from courts, agencies would be well advised to pay attention to practices in the circuit courts, as agencies are much more likely to have their rules reviewed there than in the Supreme Court.

Second, the pool of cases decided by the Supreme Court likely shows strong "selection effects"[91]: they cannot be viewed as a representative sample of the entire population of agency statutory interpretation cases. With plenary control over the certiorari power, the Supreme Court can grant review to whichever cases it wishes, and those that attract the Court's

86. Peter H. Schuck & E. Donald Elliott, *To the* Chevron *Station: An Empirical Study of Federal Administrative Law*, 1990 DUKE L.J. 984. For a fuller review of empirical studies on Supreme Court deference practice, see Raso & Eskridge, *supra* note 6, at 1739–42.

87. *See* Eskridge & Baer, *supra* note 6; Raso & Eskridge, *supra* note 6. The dataset is available in the IQSS Dataverse Network and on file with the author. *See Replication Data For: The Continuum of Deference: Supreme Court Treatment of Agency Statutory Interpretations from* Chevron *to* Hamdan, IQSS, http://dvn.iq.harvard.edu/dvn/faces/study/StudyPage.xhtml?globalId=hdl:1902.1/16562&studyListingIndex=0_354424ece633571d155824566165.

88. Eskridge & Baer, *supra* note 6, at 1094. The Raso and Eskridge article uses a 667-case subset of this dataset, consisting of all cases where the agency interpretation at issue was not advanced for the first time in the course of litigation. *See* Raso and Eskridge, *supra* note 6, at 1741.

89. For a listing and explanation of the variables, see Eskridge & Baer, *supra* note 6, at 1203–24.

90. *See* Peter L. Strauss, *One Hundred Fifty Cases Per Year: Some Implications of the Supreme Court's Limited Resources for Judicial Review of Agency Action*, 87 COLUM. L. REV. 1093, 1095 (1987) (demonstrating the infrequency with which the Supreme Court reviews lower court decisions and observing the freedom this gives to lower courts to alter existing law).

91. Eskridge & Baer, *supra* note 6, at 1096–97.

attention are likely to be an exceptional bunch. As Eskridge and Baer point out, these probably include a disproportionate share of the most difficult cases,[92] and they may be aberrant for other reasons as well.[93] As a result, we cannot assume that these cases offer a reliable guide to how the federal judiciary as a whole applies deference.

Third, although the dataset is large, the number of potentially relevant variables is large as well, meaning it is difficult to draw firm conclusions about which (if any) factors bear on the Court's practice of deference.[94] Once we drill down to see how results vary by individual agency or subject matter area, for instance, we have few cases in any given condition,[95] making it difficult to tell whether variations in outcome reflect systematic differences or random noise. When we look only at challenges to notice-and-comment regulations and formal adjudications—the focus of attention in this Article—the dearth of data is even more acute.[96]

Finally, these data provide limited leverage for answering the important question of how much, and what kind of, scrutiny the Court is applying under *Chevron* and *Skidmore*. If the *Chevron* lottery is, in fact, a lottery—that is, if agencies cannot reliably anticipate which of their statutory interpretations will be reviewed under *Chevron* as opposed to *Skidmore*—then a higher survival rate among the former is evidence that *Chevron* review is more lax than *Skidmore*, as the doctrine asserts.[97] Without a measure of the interpretive content of challenged rules, however, we cannot use these data to evaluate the features of the *Skidmore* lottery—that is, to assess how the

92. *Id.*

93. *See id.* at 1097 ("Even less predictable is the effect of the much-touted '*Chevron* Revolution' . . . on the kinds of cases that are litigated and appealed . . . at the Supreme Court level.").

94. As Eskridge and Baer point out, their dataset is not a "sample" at all, but rather the complete universe of agency interpretation cases during the time period they study. But to the extent their study provides a guide to the Court's conduct going forward, it is a time-specific sample; one would think that patterns identified here would apply also to the years since 2006.

95. *See* Eskridge & Baer, *supra* note 6, 1204–05 (displaying the results by agency and subject matter).

96. To take one example, in the entire Eskridge and Baer dataset, there are only six cases in which the statutory interpretation is advanced in a legislative rule from the Environmental Protection Agency, and not all of these were attended by notice-and-comment procedures.

97. This is, in fact, what the data show. *See infra* note 106. This finding contradicts the claim that similar reversal rates across different standards of review illustrate that the ostensibly different standards amount, at base, to a single "reasonableness" standard. *See* David Zaring, *Reasonable Agencies*, 96 VA. L. REV. 135, 137 (2010) (arguing that *Chevron* and *Skidmore* deference essentially amounts to a single "reasonableness" standard). If agencies are in a position to anticipate which standard of review will be applied, the fact that agency survival rates hover around 70% under different standards of review need not mean that the various standards are equivalently stringent. For instance, when agencies interpret statutes in guidance documents, they know they are most likely to receive *Skidmore* deference. *See* United States v. Mead Corp., 533 U.S. 218, 234 (2001) (reaffirming *Skidmore*'s holding that "an agency's interpretation may merit some deference"). A rational strategy for the agency would be to avoid more adventurous interpretations in guidance documents.

probability of surviving review under *Skidmore* varies with the content of the statutory interpretation. Fortunately, qualitative work on the courts of appeal offers some insight into this question.[98]

All that being said, Eskridge and Baer's research broadly corroborates this Article's characterization of the *Chevron* lottery. As Eskridge and Baer themselves state, "Are there factors that predict . . . when particular deference regimes will be invoked . . . ? [O]ur data offer little to latch onto; there is no clear guide as to when the Court will invoke particular deference regimes, and why."[99] This section now turns to explore what the data show us in more detail.

First, the Eskridge and Baer data establish that the interpretations that are the focus of this Article—those advanced in formal adjudications and legislative rules issued through notice-and-comment rulemaking—are more likely to get *Chevron* deference than those adopted through less formal proceedings.[100] The Court applied *Chevron*, or an equivalent or more deferential standard (*Beth Israel*[101] or *Curtiss-Wright*[102] deference, in Eskridge and Baer's terminology[103]) in 44% of the cases involving statutory interpretation through notice-and-comment rulemaking or formal adjudication.[104] This compares to an application of *Chevron* or the equivalent in only 5% of cases involving less formal interpretations. A glance at these results makes clear, though, that even for these more formal interpretations, agencies can hardly count on getting *Chevron* deference. While the doctrine suggests a strong presumption that notice-and-comment rules and formal adjudications will receive *Chevron* deference,[105] the data show a high probability that the Supreme Court may apply *Skidmore* instead.[106]

---

98. *See infra* section II(B)(2).

99. Eskridge & Baer, *supra* note 6, at 1091.

100. *Id.* at 1149 tbl.18.

101. Beth Israel Hosp. v. NLRB, 437 U.S. 483 (1978).

102. United States v. Curtiss-Wright Export Corp., 299 U.S. 304 (1936).

103. Eskridge & Baer, *supra* note 6, at 1098. Eskridge and Baer also list *Seminole Rock* deference as a standard more deferential to agencies than *Chevron*, but it is inapplicable here because it applies only to an agency's interpretations of its own regulations, not statutes. *Id.*

104. In discussing Eskridge and Baer's data below, I use the terms "apply *Chevron*" and "receive *Chevron*" as shorthands to mean that the Court approaches the statutory interpretation question using the *Chevron* framework, or these equivalent or still more deferential standards. This does not necessarily mean either that the Court finds the statutory language to be ambiguous or that the Court accepts the agency's interpretation—although the model I develop in Part III starts from the assumption that the agency wins when the Court applies *Chevron*, an assumption I later relax. The figures appearing in this subpart are my own calculations based on Eskridge and Baer's data.

105. United States v. Mead Corp., 533 U.S. 218, 230 (2001). *But see* Nat'l Cable & Telecomms. Ass'n v. Brand X Internet Servs., 545 U.S. 967, 1004 (2005) (Breyer, J., concurring) (noting that observing the more formal procedures is not sufficient to guarantee *Chevron*'s application).

106. The agency's chances of winning before the Court are lower when *Skidmore* is applied than when *Chevron* is applied. Of formal adjudications and notice-and-comment rulemakings in the

Within this population of agency interpretations, are there features of these rules that help explain whether *Chevron* or *Skidmore* can be applied? Some agencies receive *Chevron* more frequently than others,[107] and the nature of the statutory grant of power has some correlation with the standard applied,[108] but it is harder to find any factor within the agency's control that it can manipulate to adjust its odds in the *Chevron* lottery.[109] One might think that the politics of an interpretation could have some bearing on how it would be reviewed, but the data do not bear this out. Eskridge and Baer coded each agency action reviewed as liberal, conservative, or neutral/mixed.[110] An interpretation is equally likely to receive *Chevron* consideration whether it has a liberal or conservative interpretation. Conservative and liberal interpretations also survive judicial review at the same rate.[111] Ultimately, when we exclude the uncodeable or neutral interpretations, the politics of agency interpretation have no statistically significant bearing on either the deference regime or the agency's win rate.

Eskridge and Baer's data do identify one respect in which the content of agency interpretations relates to the deference standard applied; and to the extent this relationship is strong, the "*Chevron* lottery" characterization is inexact. Eskridge and Baer evaluate whether an agency interpretation is (1) long-standing and fairly stable, (2) evolving, or (3) recent.[112] I find that the Court applied the *Chevron* framework somewhat more frequently, and agencies won more often, when they maintained a long-standing and stable interpretation. The Court applied *Chevron* 48% of the time when the agency

---

Supreme Court, agencies won 78% of those to which *Chevron* was applied, as opposed to 67% of those to which *Chevron* was not applied.

107. For instance, the IRS receives *Chevron* for only 33% of its interpretations, while for the Department of Health and Human Services, the figure is 74%. These agency-based discrepancies are due in some part to subject-specific lines of doctrine that in some cases seem to be eroding. *See, e.g.*, Mayo Found. for Med. Educ. & Research v. United States, 131 S. Ct. 704, 706–07 (2011) (dropping the tax-specific *National Muffler* standard in favor of *Chevron* analysis for Treasury regulations).

108. Eskridge and Baer code for different statutory grants of authority, and the strongest, "Merrill-Watts" form of delegation—in which the agency is authorized to impose immediate sanctions for violations of its rules or orders—is associated with a higher incidence of *Chevron* deference. Eskridge & Baer, *supra* note 6, at 1125–26, 1209–10; Merrill & Watts, *supra* note 78, at 582–86 (describing a way of interpreting congressional delegations of authority). Not surprisingly, the forms of delegation also tend to vary systematically by agency. For instance the IRS lacks "Merrill-Watts" delegation, and the Department of Health and Human Services frequently proceeds under Merrill-Watts delegations of power. Eskridge & Baer, *supra* note 6, at 1210.

109. The data do suggest, in addition, that the rate at which *Chevron* is applied varies somewhat by agency, and also by whether the agencies are acting pursuant to an express delegation of legislative power, which is substantially correlated with the former. These are thus examples of variables that seem to vary systematically with the incidence of *Chevron* review, but they are not factors that agencies are in a position to control.

110. Eskridge & Baer, *supra* note 6, at 1205.

111. Conservative interpretations survive review 69.4% of the time, and liberal interpretations, 68.12% of the time.

112. Eskridge & Baer, *supra* note 6, at 1206. In the analysis performed for this Article, both of these latter two categories are recoded as representing a change in agency interpretation.

interpretation advanced was continuous, and 40% of the time when the interpretation was recent or evolving. If we ignore formal adjudications and restrict our attention to notice-and-comment rulemakings, however, the discrepancy is larger: continuous interpretations received *Chevron* review 51% of the time, compared to 35% for recent or evolving interpretations. A statistical test suggests that the difference in frequency is larger than we would expect to occur by chance.[113]

This result is surprising, and goes unremarked on by Eskridge and Baer. Under *Mead*, the continuity of an interpretation should have no bearing on the appropriateness of *Chevron* review, although *Barnhart*, by contrast, lists as one factor in the *Chevron* Step Zero analysis "careful consideration the Agency has given the question over a long period of time."[114] The discrepancy in *Chevron* application rates suggests that perhaps *Barnhart*, though infrequently cited by the Supreme Court, may play a larger role in guiding the Court's *Chevron* Step Zero analysis than generally recognized. Alternatively, it raises the intriguing possibility that, if durability of an interpretation is in some way a proxy for its "quality," the Court is manipulating its deference analysis by applying *Chevron* as a cover in cases where the Court in fact agrees with the interpretation on the merits. It falls outside the scope of this project to investigate this possibility further. In any event, whatever relevance continuity may have had to the *Chevron* Step Zero analysis for notice-and-comment rulemakings, it appears to be on the wane. The Court has moved in recent years to devalue continuity of agency practice in administrative law more generally,[115] and if we confine our attention to cases decided since the 2000 Term (when *Mead* was decided), the trend vanishes.

Moreover, it is important to note that the Eskridge–Baer data may exaggerate the unpredictability of *Chevron* Step Zero decisions for reasons having to do with the nature of common law decision making and the organization of the federal courts. Once a court has determined which deference standard governs a particular agency's interpretations of a particular statute, the issue is settled, at least for that court and those bound by its decisions. Over time, then, the set of cases subject to the *Chevron* lottery could dwindle, as each ruling on what standard governs a given

113. Specifically, a Pearson chi-squared test yields a value of 5.1654, with an associated probability of 0.023: in other words, if continuous and noncontinuous interpretations were in fact treated the same way, we would expect to see a difference this large emerge by chance only 2.3% of the time.

114. Barnhart v. Walton, 535 U.S. 212, 222 (2002); *see also supra* text accompanying note 81.

115. *See, e.g.*, FCC v. Fox Television Stations, Inc., 129 S. Ct. 1800, 1810–12 (2009) (explaining that an agency's change in position does not trigger a heightened standard of review and does not require justifications for the new policy that is any "more substantial than those required to adopt a policy in the first instance").

agency/statute pairing eliminates uncertainty going forward.[116] Also, the Eskridge–Baer data concern deference practices in the Supreme Court only. Unpredictability may be systematically higher in the Supreme Court than in federal appeals courts, either because (1) issues of first impression form a larger part of the Supreme Court's docket,[117] or (2) the Supreme Court is more inclined than lower courts to deny *Chevron* review to notice-and-comment rulemakings or formal adjudications.[118]

These concerns could diminish the aptness of the *Chevron* lottery as a characterization of the environment agencies face on judicial review. How seriously they undermine the lottery characterization is ultimately an empirical question that I lack the data to answer adequately. Still, there are reasons to think the *Chevron* lottery remains an appropriate metaphor notwithstanding these concerns. First, while judicial decisions reduce uncertainty about which standard applies in a particular context, Congress continually replenishes the supply of uncertainty by establishing new statutes and new agencies. Second, only the Supreme Court's rulings are binding on all circuit courts, and the Supreme Court itself hears few administrative law cases a term.[119] Meanwhile, a baker's dozen of circuit courts continue to review agency decisions, treating cases from sister circuits as persuasive authority only. If unpredictability is being squeezed out of the deference regime, it is being squeezed out gradually.

To sum up: the Eskridge–Baer data reveal that a significant proportion of agency statutory interpretations adopted in notice-and-comment rulemaking or formal adjudications receive *Skidmore* deference, even though *Mead* suggests a strong presumption in favor of *Chevron*.[120] The data also suggest that there is little a given agency can do to nudge its probability of getting *Chevron* up or down in any particular rulemaking.[121] The politics of the interpretation have no bearing on the deference regime at all.[122] The data

116. For instance, in the past few years, the Supreme Court has ruled that *Chevron* governs Treasury regulations interpreting the tax code (Mayo Found. for Med. Educ. & Research v. United States, 131 S. Ct. 704 (2011)), FCC regulations interpreting the Telecommunications Act (Nat'l Cable & Telecomms. Ass'n v. Brand X Internet Servs., 545 U.S. 967 (2005)), and Federal Reserve Board regulations interpreting the Truth in Lending Act (Household Credit Servs., Inc. v. Pfennig, 541 U.S. 232 (2004)). My thanks to Kristin Hickman for these examples.

117. *See, e.g.*, Stefanie A. Lindquist & Frank B. Cross, *Empirically Testing Dworkin's Chain Novel Theory: Studying the Path of Precedent*, 80 N.Y.U. L. REV. 1156, 1173 (2005) (explaining that under path dependence theory, "initial cases of first impression allow great judicial freedom").

118. *See* Eskridge & Baer, *supra* note 6, at 1119–20 (speculating that the Court views deference regimes as guides for lower courts, but is more flexible in its own decision to use such regimes).

119. *See* Strauss, *supra* note 90, at 1099 (stating that the Supreme Court hears only "a handful" of such cases each year). The Supreme Court's docket has further fallen sharply since the appearance of Strauss's article, over a quarter century ago.

120. *See supra* notes 100–06 and accompanying text.

121. *See* Eskridge & Baer, *supra* note 6, at 1137 (stating that "[t]he Court does not apply deference regimes in a foreseeable manner" but instead "invokes deference regimes in a manner that is seemingly sporadic and haphazard").

122. *See supra* notes 110–11 and accompanying text.

do suggest that an agency can increase its odds of *Chevron*, at least in the notice-and-comment rulemaking context, by maintaining continuity in its interpretation over time.[123] To the extent this observed effect is real, it is at odds with this Article's characterization of the *Chevron* lottery, an environment in which the chances of getting *Chevron* are beyond the agency's power to shape. That said, the evidence also suggests that continuity no longer matters to the Court's choice of deference regime, even if it did in the past.[124] On the whole, then, to the extent these data are revealing, they comport well with this Article's characterization of the *Chevron* lottery.

2. Deference in the Courts of Appeals.—This section turns to work on deference practice in the courts of appeals. Scholarship in this area combines quantitative and qualitative methods to shed light on how *Skidmore* is applied.

For the purposes of this Article, the most valuable source on the circuit courts is a study by Kristin Hickman and Matthew Krueger.[125] The work provides a helpful complement to the Eskridge–Baer and Razo–Eskridge pieces. Its scope is narrower: it examines a smaller set of cases, those applying *Skidmore* in circuit courts between summer 2001 and summer 2006, 106 in all.[126] But if the dataset is smaller and the focus narrower, the work provides a close, qualitative analysis. The authors read the cases to characterize the nature of deference applied under *Skidmore*, and their study is the first to examine in depth the analysis conducted in a large population of *Skidmore* cases.[127]

The authors contrast two models of *Skidmore* analysis that are rooted in the case law: a sliding-scale model, in which courts are sensitive to indicia of agencies' reliability and fidelity, and an independent-judgment model that is tantamount to de novo review.[128] Their core finding is that, generally, the circuit courts follow the sliding-scale model.[129]

---

123. Eskridge & Baer, *supra* note 6, at 1133.

124. *See supra* notes 114–16 and accompanying text.

125. Kristin E. Hickman & Matthew D. Krueger, *In Search of the Modern* Skidmore *Standard*, 107 COLUM. L. REV. 1235 (2007). Other important studies of the impact of *Mead* in the lower courts include articles by Lisa Schultz Bressman and Adrian Vermeule. Both pieces show courts struggling to find their way in the early aftermath of *Mead*. Writing in 2003, Vermeule reported that "[i]n the trenches of the D.C. Circuit, . . . *Mead*'s ambitious recasting of deference law has gone badly awry," as panels reached inconsistent, and in Vermeule's view, frequently mistaken, views of what *Mead* required. Vermeule, *supra* note 10, at 349. Writing two years later, Bressman observed courts dividing over whether to follow *Mead* or *Barnhart*. Bressman, *supra* note 5, at 1459. Note that subsequent history suggests courts' enthusiasm for *Barnhart* to be a passing fancy: court of appeals cases cited *Barnhart* twenty-eight times in 2003 alone, but only five to ten times annually in recent years.

126. Hickman & Krueger, *supra* note 125, at 1259–60.

127. *Id.* at 1267.

128. *Id.* at 1252–59.

129. *Id.* at 1271.

Hickman and Krueger's conclusions bolster this Article's characterization of the *Skidmore* lottery as an environment in which the probability of surviving judicial review is pegged to the plausibility of the agency's interpretation. Hickman and Krueger find that courts applied the sliding-scale approach to *Skidmore* in 79 of the 106 cases reviewed, and the independent judgment model in only 20.[130] They further find that, of the factors named in *Skidmore*, the two given the most emphasis by reviewing courts are the validity of the agency's reasoning and the thoroughness of its consideration.[131] Hickman and Krueger understand "validity" "to include discussion of the reasonableness and plausibility of the interpretation itself,"[132] and note that in evaluating the "thoroughness of consideration," courts examine the quality of the justification proffered by the agency.[133] Courts give comparatively less weight to other "contextual" factors: the formality of the agency's procedures,[134] the consistency of the agency's interpretation over time,[135] and the agency's subject-specific expertise.[136]

Putting this all together, when circuit courts apply *Skidmore*, they are generally applying sliding-scale deference. And in applying sliding-scale deference, they focus primarily on the content of the agency's interpretation, and specifically, its apparent consistency with the statute. Other, contextual factors that are independent of the interpretation's content—such as agency expertise and formality of procedures—turn out to play a less prominent role in *Skidmore* analysis than sometimes thought.[137] In other words, when courts apply *Skidmore* review, the chance an interpretation will survive rises to the extent that it is credible as a faithful construction of the statute.

To be sure, the stylized account this Article presents of how sliding-scale deference works in the *Skidmore* lottery does not capture the full complexity of *Skidmore* review in practice. No doubt in many instances, statutory terms may be so open-ended that there is no way to say which possible interpretation better keeps faith with the statute.[138] In some cases,

130. *Id.*

131. *Id.* at 1281, 1285.

132. *Id.* at 1285.

133. *Id.* at 1281.

134. *See id.* at 1283 ("Courts assessed the formality of the administrative interpretation's procedural pedigree and format with somewhat less frequency than other factors.").

135. *See id.* at 1286 ("[D]espite its numerous appearances in judicial opinions, 'consistency' seems less dispositive than other *Skidmore* factors.").

136. *See id.* at 1288–89 ("[T]he expertise factor generally lacks teeth, as courts only counted this factor against agency deference in three of the cases we evaluated.").

137. *See* Ronald J. Krotoszynski, Jr., *Why Deference?: Implied Delegations, Agency Expertise, and the Misplaced Legacy of* Skidmore, 54 ADMIN. L. REV. 735, 737–39 (2002) (arguing for an interpretation that emphasizes the importance of expertise).

138. *Skidmore* itself may have been such a case. Again, the question at issue was whether the nights that firefighters spent on call at the plant counted as "working time" for purposes of the Fair Labor Standards Act. *See supra* notes 48–49 and accompanying text. The traditional tools of statutory construction aided the Court little in determining which interpretation better comported with the enacting Congress's intent.

the answer may depend on what theory of statutory interpretation one endorses.[139] But though this Article's account of the *Skidmore* lottery is a simplification, it is a reasonable one. The Article's core contention is that, all else equal, the more persuasive an agency's claim that an interpretation is consistent with the statute, the better chance it will stand under *Skidmore* review. And this is consistent with Krueger and Hickman's finding that *Skidmore* is generally applied as a sliding-scale deference standard along the lines discussed above.[140]

A case such as *Lopez v. Terrell*[141] illustrates how sliding-scale *Skidmore* review works in practice. The case presented the question of whether a federal prisoner accrues "Good Conduct Time" (GCT) for time spent in federal and state custody before his federal sentencing.[142] The applicable statute provided that prisoners may receive "up to 54 days [GCT] at the end of each year of the prisoner's term of imprisonment."[143] In informal rulings, the Bureau of Prisons (BOP) had interpreted the statutory language to permit the accrual of GCT only for time spent in federal custody following sentencing.[144] Inmate Lopez had argued, and the district court had agreed, that the statutory language permitted the accrual of GCT for all the time spent in custody for his federal offense.[145]

The appellate court evaluated the BOP's interpretation under *Skidmore*.[146] Rather than interpreting the contested language for itself, or calibrating the deference owed the agency based on its expertise, the court carefully considered the agency's case for its reading of the statute. BOP made a tight textualist argument, arguing "that the phrase ['term of imprisonment'] must be understood within the context of the statute as a whole and, in particular, in reference to the word 'sentence' in the preceding phrase, 'a prisoner . . . may receive credit toward the service of the prisoner's *sentence*.'"[147] The agency argued that the meaning of "sentence" was clearly defined in federal law, showed where, and explained how that definition

139. *See* WILLIAM N. ESKRIDGE, JR. ET AL., CASES AND MATERIALS ON LEGISLATION: STATUTES AND THE CREATION OF PUBLIC POLICY 689–846 (4th ed. 2007) (exploring intentionalist, purposivist, and textualist approaches to statutory construction). The Supreme Court has described the object of statutory interpretation as determining congressional intent, "[e]mploying traditional tools of statutory construction." INS v. Cardoza-Fonseca, 480 U.S. 421, 446 (1987). This Article need not—and does not—take a position in the rich theoretical debates over precisely what this means. This is because, generally speaking, an interpretation of an ambiguous statute that satisfies *Skidmore* will tend to be truer to the statute, by the lights of *any* plausible theory of statutory interpretation, than an interpretation that fails to do so.

140. *See supra* notes 128–29 and accompanying text.

141. 654 F.3d 176 (2d Cir. 2011).

142. *Id.* at 177.

143. *Id.* (citing 18 U.S.C. § 3624(b) (2006)).

144. *Id.* at 180.

145. *Id.*

146. *Id.* at 183.

147. *Id.* (citing 18 U.S.C. § 3624(b)).

foreclosed Lopez's interpretation.[148] Noting that it "aligns with traditional canons of statutory interpretation," the court "[found] the BOP's construction of [the statute] persuasive" under *Skidmore*.[149]

To sum up this subpart: empirical evidence gathered from both Supreme Court and circuit court practice supports this Article's claim that agencies facing judicial review of their statutory interpretations face a deference lottery with specific features. Even for interpretations offered in notice-and-comment rulemakings and formal adjudications, where *Mead* suggests a presumption of *Chevron* review, Eskridge and Baer's Supreme Court data show that *Skidmore* is frequently applied instead.[150] Moreover, the deference regime applied in individual cases is not correlated with objective measures of an interpretation's content,[151] meaning that, from the agency's perspective, the standard applied seems to be chosen as if through a random draw. This is consistent with the Article's characterization of the *Chevron* lottery. When *Skidmore* is applied, Hickman and Krueger's study of circuit courts shows that panels are attentive to the plausibility of agencies' statutory interpretation, more so than to content-independent contextual factors such as agency expertise.[152] This finding tracks my account of the *Skidmore* lottery, in which agencies cannot ensure their survival under *Skidmore* review, but can improve their chances by choosing an interpretation that is safer—that is, more credibly faithful to the statute.

C. *Why a Deference Lottery?*

This Part has established that agencies face a deference lottery when they defend statutory interpretations in court. The ultimate object of this Article is not to explain *why* agencies encounter a deference lottery, but to explore *how* agencies would rationally respond to one.[153] In other words, the focus of this Article is squarely on the behavior of agencies in reaction to deference lotteries, not the behavior of courts that gives rise to them. That being said, this subpart very briefly explains why the existence of a deference lottery is in fact consistent with some common-sensical suppositions about judicial behavior. The subpart concludes with a simple illustration of how the *Skidmore* lottery described in this Article can arise as the product of individual judicial decisions.

The deference lottery concept implies that the outcomes we are interested in—whether an agency gets *Chevron* or *Skidmore* deference (the

148. *Id.*

149. *Id.*

150. *See supra* text accompanying notes 104–05.

151. As noted above, there is an exception for notice-and-comment rulemakings prior to 2000, where the probability of *Chevron* review is somewhat higher when interpretations are long-standing or continuous. *See supra* text accompanying notes 112–15.

152. *See supra* text accompanying notes 129–36.

153. *See infra* Part III.

Chevron lottery) and whether or not, under *Skidmore*, the agency is upheld (the *Skidmore* lottery)—are selected as if through random draws from given probability distributions. For this to be so, it need not be the case that individual judges are actually randomizing their decisions: flipping coins to decide cases, as it were. Rather, so long as individual judges differ sufficiently in how they apply the substantive standards, it is enough that the *assignment* of judges to panels be random.[154] The deference lottery is not an attribute of the practice of any individual judge, but is instead an *emergent feature* of the judicial system.[155]

Consider, for instance, the *Chevron* lottery that governs whether agency statutory interpretations will be evaluated under *Chevron* or *Skidmore*. As a matter of doctrine, this decision is governed by the *Mead* standard: did Congress intend for the agency to speak to this issue with the force of law? But as discussed,[156] the *Mead* standard is so vague that we can expect individual judges to differ over how liberally *Chevron* deference should be granted under it. Assuming that judges vary in their disposition to apply *Chevron* across a court, three-judge panels drawn at random from that court will reach different conclusions. The probability that a randomly selected panel will apply *Chevron* is a function of the distribution of views about *Chevron*'s scope in the pool of judges. The net effect is to randomize what standard is applied, although no individual judge is randomizing.

A similar story explains the emergence of the *Skidmore* lottery. Under *Skidmore*, an agency interpretation merits deference proportional to its "power to persuade."[157] This, too, is a vague standard, and judges may differ on just how "persuasive" an agency interpretation must be to survive judicial review under *Skidmore*. Depending on which judges are selected to hear a case, a given interpretation may or may not pass muster. But the more clearly faithful to the statute an interpretation is, the better chances it has of surviving *Skidmore* review, because a larger number of possible panels may deem it acceptable.

This point can be developed more formally. Suppose that the different possible interpretations of a statute are laid out on a continuum, from the

---

154. On randomization in judicial assignment to appellate court panels, and divergences from strict randomness, see Matthew Hall, *Randomness Reconsidered: Modeling Random Judicial Assignment in the U.S. Courts of Appeals*, 7 J. EMPIRICAL LEGAL STUD. 574, 577–81 (2010).

Indeed, the assignment of judges need not even be random for judicial review to approximate a lottery, so long as individual judges' views on how the standards apply in concrete cases are sufficiently opaque from the agency's perspective. The Supreme Court, of course, hears its cases en banc, and yet it is no easy feat predicting how the Court will apply its deference regime to particular cases, as Eskridge and Baer show. Eskridge & Baer, *supra* note 6, at 1136–53.

155. "Emergence" refers to the development of complex systems through the aggregation of simple individual behaviors. *See generally* STEVEN JOHNSON, EMERGENCE: THE CONNECTED LIVES OF ANTS, BRAINS, CITIES, AND SOFTWARE (2001) (chronicling various systems displaying emergent features).

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

157. Skidmore v. Swift & Co., 323 U.S. 134, 140 (1944).

most defensible (i.e., most readily justified as faithful) to the most adventurous. Imagine that each judge on a court has a "decision cutpoint" falling somewhere along this line: in his view, no interpretations more adventurous than this cutpoint are justifiable under *Skidmore* review. Imagine that this court has ten judges, whose cutpoints are distributed as shown in Figure 1. The point 0 on the axis denotes an interpretation so extreme that no judge would approve it under *Skidmore*. Points 0.1 through 1 each represent the cutpoint of one of the ten judges.

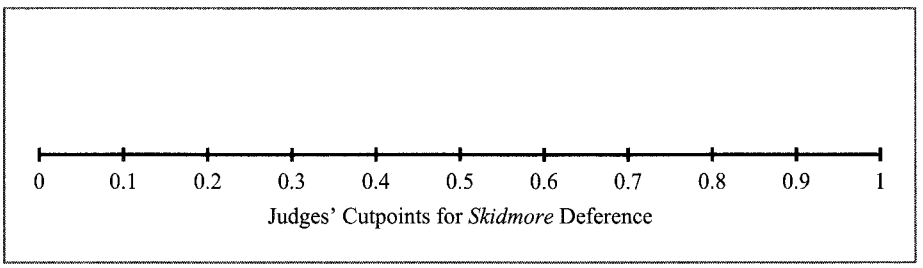

Figure 1: Distribution of Cutpoints.

A panel of three judges will be drawn at random from this court to evaluate an agency's statutory interpretation. In the event that the court applies *Skidmore*, how does the outcome of judicial review vary as a function of which interpretation along this continuum it chooses? Figure 2 illustrates the answer. The court's judgment varies probabilistically as a function of the interpretation the agency chooses. Moving along the continuum from left to right, the interpretation falls to the right side of more and more judges' cutpoints, so that the chances rise of drawing a panel of three judges, of whom two would approve it. The actual shape of the probability distribution will depend on the distribution of judges' cutpoints along the line.[158] Obviously, this stylized representation greatly simplifies the actual practice of deference and decision making on appellate courts.[159] But it suffices to show that a deference lottery follows naturally from reasonable assumptions about how multi-member courts operating in panels apply vague standards.

158. The calculations of the probabilities shown in the graph are on file with the author and available upon request.

159. For a more detailed, formal treatment of judicial bargaining on three-judge courts, see Jud Mathews, *Opinion Competition and Judge Replacement on Collegial Courts* (Ill. Program in Law, Behavior, & Soc. Sci., Paper No. LBSS12-19, 2012), *available at* http://ssrn.com/abstract=1868619.

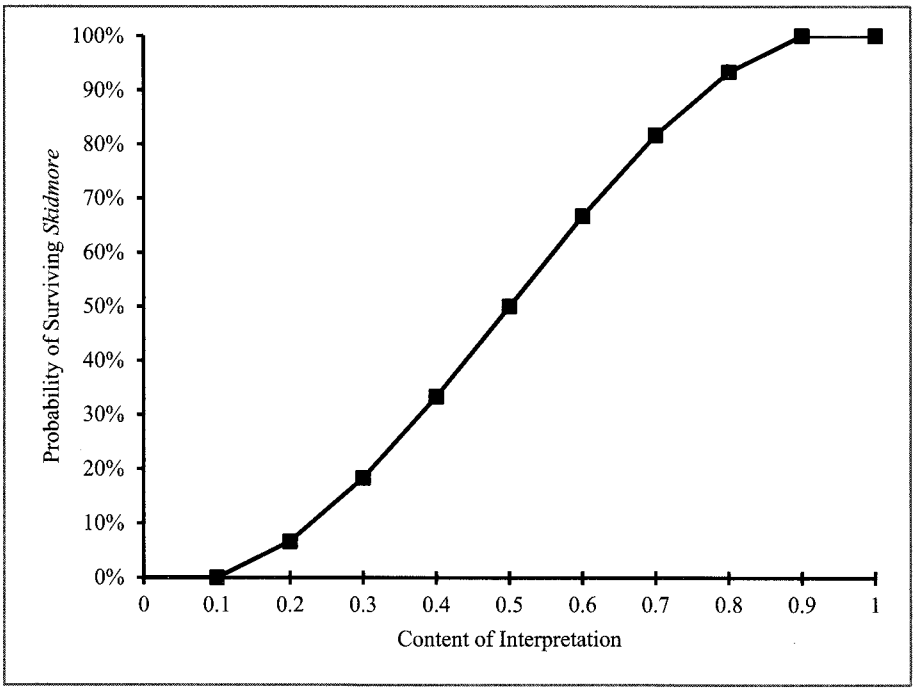

Figure 2: Skidmore *Lottery as Aggregation of Judges' Cutpoints.*

Note that this account is entirely consistent with strategic behavior by judges in pursuit of their own policy preferences.[160] Individual judges may be more disposed to defer to conservative rulings or liberal ones, and more or less willing to stretch doctrine in pursuit of their preferences.[161] This framework can easily accommodate judges with policy preferences if we suppose that judges' preferences reflect where their decision cutpoints lie in individual cases. For the deference lottery to work, doctrine need only provide at least a weak constraint on judicial action—preventing, for instance, judges from having "backwards" cutpoints, so that *more* extreme interpretations are *less* likely to be rejected. The other judges on a panel are in a position to check, or at the very least, spotlight politically motivated rulings that are sharply at odds with doctrine—for instance, rulings against

160. For empirical work supporting the thesis that judges manipulate review of agencies to achieve favored results under hard look review, see Thomas J. Miles & Cass R. Sunstein, *The Real World of Arbitrariness Review*, 75 U. CHI. L. REV. 761 (2008).

161. For the classic statement of the "attitudinal" model of judicial decision making, see JEFFREY A. SEGAL & HAROLD J. SPAETH, THE SUPREME COURT AND THE ATTITUDINAL MODEL REVISITED 86–97 (2002).

the agency when its interpretation is plainly justifiable under the *Skidmore* formula.[162]

III. Playing the Deference Lottery

Having established the deference lottery's existence in the previous Part, this Part considers its implications for agency behavior. In particular, it investigates how different configurations of the deference regime's component parts—the *Chevron* lottery and the *Skidmore* lottery, in the parlance of this Article—might induce different kinds of statutory interpretations on the part of agencies. The method for exploring these dynamics come from the tool kit of Positive Political Theory (PPT): a simple, decision-theoretic model of agency action.[163] This Part lays out the model and highlights some of its key results. Part IV will consider what practical implications this deductive exercise has for administrative law and courts' deployment of deference doctrines.

As noted in Part I,[164] this approach to the problem of deference is strongly influenced by PA theory.[165] From the perspective of this model, agencies are regarded as agents—first and foremost, of the Congresses that enacted the statutes they administer, or (more abstractly) of the statutes themselves. When Congress charges an agency with administering a statute, Congress intends the agency to carry out its statutory charge faithfully. However, Congress's delegation of authority to the agency introduces slack and creates the possibility for "agency losses"—for the agency to pursue its own ends, rather than the principal's.[166] Statutory interpretation is one means by which agencies can slant the administration of a statute to the service of their own policy priorities. For instance, in the 1970s, the National Labor Relations Board sought to broaden the definition of "employee" to include

---

162. *See* Jonathan P. Kastellec, *Panel Composition and Judicial Compliance on the U.S. Courts of Appeals*, 23 J.L. ECON. & ORG. 421, 425–27 (2007) (discussing how dissents and the threat of dissents can act as a constraint on majority rulings that stray too far from established doctrine).

163. The essence of PPT is to "treat[] policymaking as a game of strategy and focus[] on the choices that rational actors make in pursuit of their goals." David S. Law, *Introduction: Positive Political Theory and the Law*, 15 J. CONTEMP. LEGAL ISSUES 1, 1 (2006). For a detailed survey of PPT's contributions to the study of public law, see McNollGast, *The Political Economy of Law*, *in* 2 HANDBOOK OF LAW AND ECONOMICS 1651 (A. Mitchell Polinsky & Steven Shavell eds., 2007).

164. *See supra* text accompanying note 24.

165. DAVID EPSTEIN & SHARYN O'HALLORAN, DELEGATING POWERS: A TRANSACTION COST POLITICS APPROACH TO POLICY MAKING UNDER SEPARATE POWERS 27–28 (1999); *see also* Mark Thatcher & Alec Stone Sweet, *Theory and Practice of Delegation to Non-Majoritarian Institutions*, *in* THE POLITICS OF DELEGATION 1, 3–9 (Mark Thatcher & Alec Stone Sweet eds., 2003) (laying out the elements of PA theory).

166. *See* EPSTEIN & O'HALLORAN, *supra* note 165, at 29 (noting that agency loss occurs when an agent generates outcomes at odds with the preferred interests of a principal).

some supervisors, despite the Taft-Hartley Act's exclusion of managerial employees from its scope.[167]

Within this PA perspective, judicial review functions as a strategy to monitor and discipline agency performance. More aggressive monitoring of agencies by courts can reduce agencies' waywardness, by inducing them to follow the statutes they administer more faithfully than they otherwise might. The idea that judicial review can have this effect is commonplace in classic administrative law theory,[168] as well as in PA scholarship.[169] However, monitoring imposes costs of its own. In particular, aggressive judicial review of agency interpretations translates into higher reversal rates: all else equal, when courts review agencies more stringently, they will strike down agency actions more frequently.[170] Reversals of agencies are extremely costly, in that they can send agencies back to the drawing board, wiping out years of work formulating a policy, and disrupt expectations among those affected by agency policy.[171]

The PA account sketched here so far fails to account for a key player in the administrative process: the President. Of course, as instrumentalities of the Executive Branch, agencies are in an important sense also agents of the President. The President is equipped with powerful tools for shaping and monitoring agency performance, starting with the constitutional authority to appoint agency officials,[172] and including the power to review agency agendas and policies on an ongoing basis through the Office of Management and Budget (OMB).[173] "Common agency" problems—where a single agent

---

167. *See* NLRB v. Yeshiva Univ., 444 U.S. 672, 674, 678 (1980); NLRB v. Bell Aerospace Co., 416 U.S. 267, 270–71, 275 (1974). On the legislative history of the Taft-Hartley Act, see Marley S. Weiss, Kentucky River *at the Intersection of Professional and Supervisory Status—Fertile Delta or Bermuda Triangle?*, *in* LABOR LAW STORIES 353, 363–64 (Laura J. Cooper & Catherine L. Fisk eds., 2005). For a discussion of contrary perspectives on the Taft-Hartley Act, see Catherine L. Fisk & Deborah C. Malamud, *The NLRB in Administrative Law Exile: Problems with Its Structure and Function and Suggestions for Reform*, 58 DUKE L.J. 2013, 2034–35 (2009). For a detailed account of changes in direction at two agencies over the course of three presidential administrations, see RICHARD A. HARRIS & SIDNEY M. MILKIS, THE POLITICS OF REGULATORY CHANGE: A TALE OF TWO AGENCIES 3–8 (2d ed. 1996).

168. *See* LOUIS L. JAFFE, JUDICIAL CONTROL OF ADMINISTRATIVE ACTION 320 (1965) (explaining the need for restrictions on agencies).

169. EPSTEIN & O'HALLORAN, *supra* note 165, at 25; Charles R. Shipan, *The Legislative Design of Judicial Review: A Formal Analysis*, 12 J. THEORETICAL POL. 269, 269 (2000).

170. Of course, a key point of this Article is that all else is *not* equal: if agencies anticipate more aggressive judicial review, they will adapt their interpretive practices strategically. But the net effect will still be a rise in reversal rates.

171. *See* JERRY L. MASHAW & DAVID L. HARFST, THE STRUGGLE FOR AUTO SAFETY 87–100 (1990) (describing in detail the costs, in terms of wasted agency resources and reduced auto safety, of judicial invalidations of vehicle safety standards).

172. U.S. CONST. art. II, § 2, cl. 2.

173. *See generally* Steven Croley, *White House Review of Agency Rulemaking: An Empirical Investigation*, 70 U. CHI. L. REV. 821 (2003) (analyzing the effects of White House review of agency rules).

has two or more principals—present their own dynamics,[174] but the President's role can also be accommodated more simply within the framework outlined here. From the perspective of the enacting Congress, the President can be regarded chiefly as a potential cause of agency losses. In other words, the President exacerbates the agency problem vis-à-vis the enacting Congress to the extent the agency is responsive to the President's agenda at the expense of a faithful construction of the statute it is charged to administer.[175]

That said, the President's responsibility "[to] take Care that the Laws be faithfully executed"[176] does bear on the model in the following important way. Focusing on the PA relationship between Congress and agency, as this Article does, is not to deny that a presidential role in steering agency policy making is appropriate or desirable. In fact, the thrust of much administrative law scholarship over the past quarter century has been to emphasize the benefits of active presidential management of agency decision making, not least because this allows for more informed and responsive policies.[177] Accordingly, the Article does not assume that the socially optimal outcome is the elimination of all agency slack, such that agencies should have no interpretive leeway. Some measure of slack may best accommodate the competing, legitimate claims of the Legislative and Executive Branches to influence the content of statutory interpretation. This Article takes no position on just how much interpretive leeway is best left to agencies, a question that is impossible to answer with any sort of precision. Rather, this Article shows that the deference lottery makes it possible for courts to elicit a wider range of interpretive behaviors from agencies than would a *Chevron*-only regime.[178] In other words, the deference lottery is a more flexible tool

174. *See* B. Douglas Bernheim & Michael D. Whinston, *Common Agency*, 54 ECONOMETRICA 923, 923 (1986) (identifying and explaining the common agency problem).

175. *See id.* at 924 (explaining that common agency problems can arise when two different government bodies oversee one agent).

176. U.S. CONST. art. II, § 3.

177. *See, e.g.*, Croley, *supra* note 173, at 821–24 (arguing for presidential review of agencies); Elena Kagan, *Presidential Administration*, 114 HARV. L. REV. 2245, 2246 (2001) (discussing the emergent role of presidential administrative control). For an overview of the presidential turn in administrative law scholarship, see Kathryn A. Watts, *Proposing a Place for Politics in Arbitrary and Capricious Review*, 119 YALE L.J. 2, 32–39 (2009).

Furthermore, as noted above, statutes are often sufficiently open-ended that there is no discernible "congressional intent" on the issues that come before agencies, so that responsiveness to the President need not come at the expense of fidelity to Congress. *See* Kagan, *supra*, at 2255–59 (reflecting on the practical extent of congressional oversight due to open-ended statutes).

Perhaps the most sustained and detailed empirical account of the systematic problems that aggressive judicial review can cause in a sensitive policy domain remains Shep Melnick's study of the Clean Air Act in the courts. R. SHEP MELNICK, REGULATION AND THE COURTS: THE CASE OF THE CLEAN AIR ACT 343–93 (1983).

178. The deference lottery is also more flexible than a *Skidmore*-only regime, so long as the inherent fuzziness of the scope-of-review language and the variation in how different judges would apply any single sliding-scale review standard combine to make a *Skidmore*-only regime, where the level of scrutiny applied is precisely calculated to produce a desired level of agency compliance, an

for shaping agency behavior, subject to the caveats and qualifications discussed below.

A. Agency Behavior and Lotteries

How do agencies decide what interpretation of a statute to advance in a regulation? Treating agencies as unitary, rational actors,[179] I assume that an agency will choose the interpretation it believes to have the highest expected value to the agency.[180] What gives an interpretation value from an agency's perspective? The model posits that agencies have policy preferences, which may diverge from the aims encoded in the statutes that they administer. If we imagine the spectrum of policy possibilities as a line segment, and the agency's most preferred policy as a point on that line (the agency's "ideal point"), I assume that, the closer a policy is to the agency's ideal point, the more value it has to the agency.[181] The agency's ideal point need not coincide with the interpretation that is most faithful to statutory intent. I also assume the agency wishes to avoid reversals by the reviewing court, and considers a reversal—in which case *no* policy takes effect—to be at least as bad as ending up with any point on the policy spectrum. For simplicity, we can think of this outcome as having a value to the agency of zero.

These assumptions are, of course, simplifications. But I argue they are reasonable first-cut approximations, suitable for this approach. Does it make sense to consider agencies to be rational, in the sense that they respond strategically to cues from courts? Close studies of agency decision making, and first-hand accounts from participants, strongly suggest that agencies do care about how their actions fare in the courts;[182] that they seek to craft agency actions to resist reversal;[183] and that they are aware, at least in broad strokes, about what standards courts are applying.[184] And while agencies are

---

unrealistic option. *See infra* text accompanying note 217. The Article focuses on *Chevron* because, for notice-and-comment rulemakings and formal adjudications, a *Chevron*-only regime seems to be the main alternative on the table. *See* United States v. Mead Corp., 533 U.S. 218, 240–41 (2001) (Scalia, J., dissenting) (arguing for *Chevron* when there is an authoritative agency interpretation).

179. *See* MCCARTY & MEIROWITZ, *supra* note 18, at 6–7 (defining rationality in game theory).

180. For a thorough discussion of subjective expected utility theory and the assumptions that underlie it, see generally PAUL ANAND, FOUNDATIONS OF RATIONAL CHOICE UNDER RISK (1993).

181. In the language of subjective utility theory, the agency's preferences are single-peaked (i.e., the ideal point is the unique maximizer for the agency) and symmetric (i.e., deviations the same distance from the ideal point to either side of it have the same value to the agency).

182. *See* Schuck & Elliott, *supra* note 86, at 1047 (showing that in 40% of remand cases, rules undergo "major changes").

183. *See* Mashaw, *supra* note 14, at 203 (describing the phenomenon of "defensive rulemaking").

184. *See* William F. Pedersen, Jr., *Formal Records and Informal Rulemaking*, 85 YALE L.J. 38, 59–60 (1975). On the other hand, there is a good argument that we need not worry that introducing the possibility of *Chevron* deference for informal agency interpretations that otherwise might merit *Skidmore* deference will induce agencies to take greater license in interpreting the statutes. The mass of such informal actions are taken on the lower rungs of agency hierarchies and tend to have limited policy salience. *See, e.g.*, United States v. Mead Corp., 533 U.S. 218, 233 (2001)

far from unitary—they are complex organizations whose component parts pursue independent, and sometimes conflicting, agendas[185]—if the concern about judicial reversal is shared within an agency, it can induce an agency to act in a way that approximates a unitary, rational actor.[186] Lastly, it is clear that agencies may act on preferences different from those encoded in the statutes they administer. Administrative law scholars and political scientists have identified many cases where agencies push policies that strain against a statutory frame, whether owing to "capture" by a set of powerful interests,[187] or issue-driven civil servants,[188] or at the behest of an administration and its political appointees.[189]

How does the agency select an interpretation? In the absence of judicial review, the agency would simply pick its ideal point: the policy that maximizes its benefit.[190] With judicial review, however, the agency has to make its selection with an eye to the rule's chances of making it past the court.[191] To put the point more formally, we can represent the expected value

---

(describing the 10,000 letter rulings issued by forty-six customs offices). There is little reason to suppose that the front-line officials issuing such interpretations are closely attuned to judicial doctrine, or that there would be a substantial difference in terms of the policy substance of their output if they were.

185. For an insightful discussion of how administrative law doctrines empower different constituencies within agencies, see Elizabeth Magill & Adrian Vermeule, *Allocating Power Within Agencies*, 120 YALE L.J. 1032, 1079–81 (2011).

186. *But see id.*; MELNICK, *supra* note 177, at 302–03 (arguing that judicial scrutiny of the Clean Air Act inflated lawyers' leverage over EPA rulemaking at the expense of agency engineers). The unitariness assumption is, however, shared by other game theoretic works on agency decision making. *See* Shipan, *supra* note 169, at 274–76 (representing the agency as a single actor in his game-theoretic model of the legislative choice of judicial review). For a selection of influential scholarship on the determinants of agency behavior, see PETER H. SCHUCK, FOUNDATIONS OF ADMINISTRATIVE LAW 82–129 (2d ed. 2004).

187. *See generally* MARVER H. BERNSTEIN, REGULATING BUSINESS BY INDEPENDENT COMMISSION (1955) (describing how the policies of the Interstate Commerce Commission at that time were dominated by railroad industry interests so that the agency no longer effectively regulated other transportation industries).

188. For a historical perspective, see, for example, DANIEL P. CARPENTER, THE FORGING OF BUREAUCRATIC AUTONOMY 131–35 (2001) (describing how a "mezzo-level" manager within the Post Office Department was integral in getting a hesitant Congress to enact permanent authority for Rural Free Delivery).

189. *See, e.g.*, Heidi Kitrosser, *Scientific Integrity: The Perils and Promise of White House Administration*, 79 FORDHAM L. REV. 2395, 2406 (2011) (illustrating Executive control over scientific information with reference to the Bush-era NASA policy of requiring all scientists' press appearances to be first cleared with the agency's public affairs office).

190. In reality, an agency would also need to steer clear of interpretations so unpalatable to the current Congress that they would elicit a statutory override, or other forms of congressional discipline. *See* EPSTEIN & O'HALLORAN, *supra* note 165, at 24–25 (describing direct and indirect methods of congressional discipline, including curtailing agency budgets).

191. Yehonatan Givati raises an intriguing possibility not explored here: that an agency might be able to choose an interpretation palatable enough to the interested parties to avoid a court challenge, and thereby short-circuit judicial review. Yehonatan Givati, *Strategic Statutory Interpretation by Administrative Agencies*, 12 AM. L. & ECON. REV. 95, 96 (2010). While this might be possible in some instances, it would probably be rare, at least with respect to important policy issues, that an agency interpretation would satisfy all potential challengers. Indeed, previous

to an agency of advancing interpretation x as $p * u_a(x)$, where p is the probability that x will survive judicial review, and $u_a(x)$ is the utility to the agency of having interpretation x become law.[192]

As I have argued, the probability that the agency's interpretation survives review—p, in the expression above—depends on both the *Chevron* lottery and the *Skidmore* lottery, and we can rewrite the expression above to take account of how each introduces unpredictability of a particular sort. I use p_c to represent the *ex ante* probability that an interpretation will receive *Chevron* review. I initially make (and later relax) the assumption that if an agency gets *Chevron*, it is home free: that the agency's interpretation will be upheld.[193] Consistent with the characterization of the *Chevron* lottery above, I also assume that p_c is independent of the content of the agency's interpretation: that the agency cannot game the odds of getting *Chevron* deference in a predictable way by manipulating the content of its rule.[194] This, then, is one "lever" the courts have over the deference lottery: how strong is the default norm that statutory interpretations are afforded *Chevron* review?

The probability that the interpretation survives sliding-scale *Skidmore* review is given by $p_s(x)$. In contrast to p_c, $p_s(x)$ depends on the content of the interpretation the agency offers: in other words, it is a function of x, which is why $p_s(x)$ is written, instead of just p_s. The model posits that the closer the agency's interpretation is to the "best" reading of the statute—the one that can most persuasively be argued to be consistent with the statute—the greater the probability it will survive *Skidmore* review. The crucial question about the *Skidmore* lottery is: do incremental shifts in the agency's interpretation shift the odds modestly or dramatically? Is *Skidmore* fairly deferential, in which case fairly adventurous readings of the statute stand a solid chance of

---

work has found that 85% of the EPA's nonroutine rules, along with every new health standard issued by OSHA has been challenged in court. CORNELIUS M. KERWIN, RULEMAKING: HOW GOVERNMENT AGENCIES WRITE LAW AND MAKE POLICY 246 (2d ed. 1999). However, the frequency of rulemaking challenges likely varies by subject matter. *See* Wendy Wagner, *Revisiting the Impact of Judicial Review on Agency Rulemakings: An Empirical Investigation*, 53 WM. & MARY L. REV. 1717, 1717 (2012) (concluding that the EPA's air toxic emissions rules are only rarely challenged in court).

192. We could also write the expected utility as $p * u_a(x) + (1 - p) * u_a(\emptyset)$, where the second term is the probability the interpretation does not survive review ($(1 - p)$) times the agency's utility from that outcome ($u_a(\emptyset)$), but the value of this term is zero, since I have stipulated that $u_a(\emptyset) = 0$.

193. Of course, agencies do not always win under *Chevron*. Still, agency interpretations stand good odds of being upheld when courts apply *Chevron*. As noted above, agencies go on to win in 78% of the cases in which the Supreme Court determines the *Chevron* framework applies. *See supra* note 106; *see also* Orin S. Kerr, *Shedding Light on* Chevron*: An Empirical Study of the* Chevron *Doctrine in the U.S. Courts of Appeals*, 15 YALE J. ON REG. 1, 31 (1998) (finding that agencies win 89% of the time when courts reach the "reasonableness" prong of the *Chevron* test).

194. As noted above, the one significant exception to this assumption suggested by the Supreme Court data is that the Supreme Court is significantly more likely to afford *Chevron* deference to consistent, rather than novel, agency interpretations. Eskridge & Baer, *supra* note 6, at 1133. But as also noted above, this regularity appears to have diminished over time. *See supra* notes 115–16 and accompanying text.

surviving review, or is it quite strict, so that agencies must hew quite close to the "best" reading to enjoy good odds? To put the point differently, if we think about increasing the odds under judicial review as buying "insurance" for the agencies against the risk of reversal, how expensive is the insurance in policy terms?[195]

The figure below illustrates the issue graphically, in a simplified form. The x-axis denotes possible interpretations of the statute, with point $x = 1$ indicating the one that can be most persuasively justified. The point $x = 0$ represents an interpretation of the statute so adventurous that it stands no chance of surviving review under any of these standards.

On the y-axis is the agency's probability of surviving *Skidmore* review. The three curves on the graph represent three possible modes of *Skidmore* analysis. The dashed line depicts a strict *Skidmore*, where movement along the x-axis towards 1 is rewarded grudgingly, with small improvements in the probability of survival; the dotted line represents a lax *Skidmore*, and the solid line, a middle-of-the-road approach.[196] As noted above,[197] there has been a long-running debate about just what level of scrutiny *Skidmore* entails, with different courts choosing among these different approaches.[198]

195. Peter Strauss's characterization of deference doctrine in terms of "*Chevron* space" and "*Skidmore* weight" identifies the same salient distinctions between the two regimes as my deference lottery concept. *Chevron* creates a zone in which agencies have the discretion to set policy themselves, whereas *Skidmore* instructs courts how much credence to give agency views in their own resolution of statutory questions. Peter L. Strauss, *"Deference" Is Too Confusing—Let's Call Them "*Chevron *Space" and "*Skidmore *Weight,"* 112 COLUM. L. REV. 1143, 1143 (2012).

196. Note that there are many other ways these curves could be drawn; the Article's only assumption is that the function is "increasing in x": in other words, that the probability of surviving *Skidmore* review goes up as the interpretation nears $x = 1$ (i.e., grows safer).

197. *See supra* notes 125–29 and accompanying text.

198. *See* Hickman & Krueger, *supra* note 125, at 1267–71 (highlighting several cases that illustrate how different courts of appeals have chosen different approaches to *Skidmore* deference—the independent judgment and sliding-scale models).

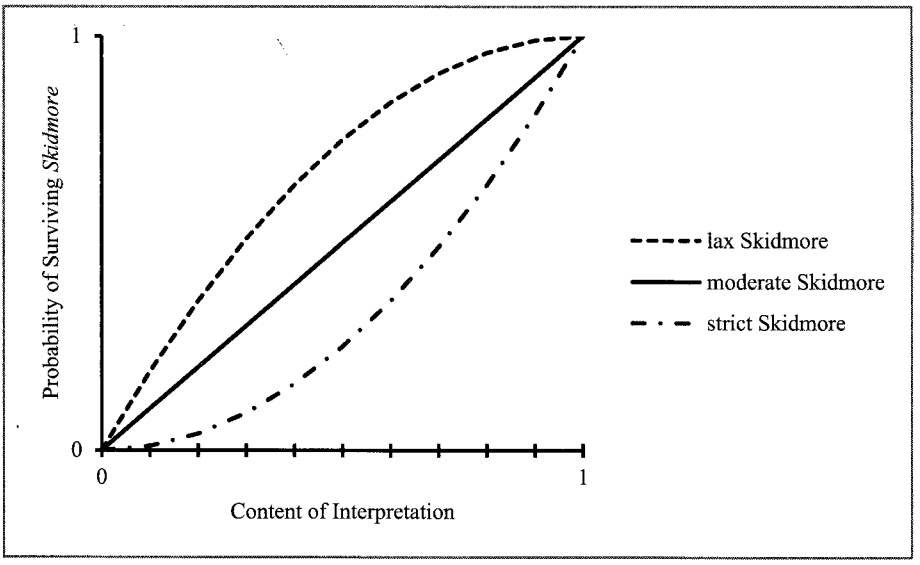

Figure 3: Three Variants of Skidmore *Review.*

Putting the pieces of the deference lottery together gives us the agency's optimization problem. The agency will select the interpretation that yields the highest expected value in light of the deference lottery—in light of the chances that it will receive sliding-scale deference instead of *Chevron*, and that it will not survive that scrutiny. Mathematically, this means choosing the value of x that maximizes this expression:

$$p_c * u_a(x) + (1 - p_c) * p_s(x) * u_a(x).$$

This is the probability of receiving *Chevron* review (p_c), times the benefit to the agency from interpretation x ($u_a(x)$), plus the probability of receiving *Skidmore* review ($1 - p_c$), times the probability that interpretation x would survive *Skidmore* ($p_s(x)$), times the benefit to the agency from interpretation x ($u_a(x)$).

B. Results

What can be said about how the deference lottery can shape agency behavior and policy outcomes in this stylized model? Of course, specific results would depend on the particulars of how an agency's utility function and preferences are defined, matters on which this Article takes no position.[199] Although the Article does not work through the agency's optimization problem formally, this subpart highlights three general results, all of which can be explained informally.

199. The results can hold whether the agency is risk neutral or risk averse, but they are more pronounced if the agency is risk averse.

1. Increasing the Stringency of Skidmore *Review Constrains Agency Opportunism—Up to a Point.*—The first observation is important, and straightforward. Relative to a *Chevron*-only deference regime, introducing the possibility that agencies' interpretations may be selected, as if at random, to face higher scrutiny can constrain agency opportunism. The chance of receiving more stringent review gives agencies an incentive to "play it safer" when interpreting statutes than they otherwise might. The mechanism comes straight from the logic of PA theory: more aggressive monitoring of agent behavior can reduce agency slack.[200] What is perhaps most noteworthy about the deference lottery is how it increases the flexibility of judicial review as a tool for managing agency behavior, relative to an across-the-board *Chevron* standard. As noted above, this Article takes no position on what is the optimal amount of agency interpretive leeway. What is notable about the deference lottery is that it can be used to incentivize *different amounts* of agency leeway, depending on how high the probability of getting *Skidmore* is and how much scrutiny *Skidmore* entails.[201] To put the point more technically, the model can yield different equilibria, depending on how the lottery is configured. This means that the deference lottery is a sensitive instrument for regulating agency conduct. Unless the desired level of agency leeway is the maximum afforded under *Chevron*, the unpredictability of the *Mead* regime, long derided as a bug, may in fact be a feature.

That being said, the model also shows that a poorly designed deference lottery can backfire. In particular, ratcheting up the scrutiny under *Skidmore* too far can have the counterintuitive—and undesirable—effect of encouraging more, rather than less, agency opportunism. One might suppose that increasing the stringency of review under *Skidmore*—that is, decreasing the deference owed to agency constructions—would always induce agencies to "play it safer" when interpreting statutes. And if *Skidmore* sliding-scale review were the only standard in play, then this would be the result.

But matters become more complicated, and more interesting, if agencies do not know *ex ante* whether they will be subject to *Chevron* review or to *Skidmore*. When *Chevron* is a possibility, increasing the stringency of *Skidmore* review past a certain point may cause a rational agency to promulgate its most preferred interpretation, rather than one calculated to win over the court with its fidelity to the statute. In technical terms, we can say that the stringency of *Skidmore* review has a "nonmonotonic" relationship to agency costs.[202] In plainer language, judicial scrutiny can backfire, leading an agency to abandon its efforts to satisfy a demanding court.

---

200. *See* EPSTEIN & O'HALLORAN, *supra* note 165, at 27–28 (discussing PA models of oversight).

201. *See supra* subpart III(A).

202. Depending on how the agency's utility function and the *Skidmore* lottery are defined, it is possible that agency may shift its interpretation towards its ideal point gradually as *Skidmore* scrutiny increases, rather than all at once after some "tipping point" of *Skidmore* scrutiny has been reached.

It is not difficult to understand why this is so. Start by imagining a deference regime in which an agency is guaranteed *Chevron* review for its regulation. Assuming—as we do to start—that the agency's most preferred interpretation would reliably be deemed "reasonable," even if it is not the most natural reading of the statute, the agency's strategy is clear: it will choose that most preferred interpretation. Now imagine that the agency faces a *Chevron* lottery: the agency has some probability of receiving *Chevron* and some probability of facing sliding-scale *Skidmore* review. The agency's strategy, as before, will be to choose the interpretation that yields the highest expected benefit to the agency. Now, however, the agency's best move in many instances will be to hedge: to pick a somewhat safer interpretation that stands a good chance of surviving *Skidmore* review, in the event that *Skidmore* is applied, but whose policy content is still satisfactory to the agency. This is the effect we want judicial review to have from a PA perspective: it induces changes in agency behavior to mitigate agency losses without generating wide-scale judicial reversals.

Imagine now the same scenario, except that the *Skidmore* standard is more strict. In other words, the "insurance" against judicial reversal in the event of *Skidmore* review has become more expensive in policy content terms: the agency must hew closer to the safest interpretation to enjoy the same probability in surviving review. At a certain point, however, the game is no longer worth the candle: the agency has a higher expected payoff from sticking with its most preferred interpretation and hoping for *Chevron*, rather than making the policy compromises needed to gain good odds of satisfying *Skidmore*. This is a bad outcome on any measure: the agency's behavior is the same as we would see in a *Chevron*-only regime, but the level of judicial reversals is higher, because the court is applying *Skidmore* some of the time.

A concrete example helps to illustrate the point. Imagine an agency is choosing between three different possible interpretations of a statute: A, B, and C, with its preferences in that order. More specifically, let's stipulate that the agency gets a benefit of 100 if interpretation A becomes law, 80 if B becomes law, and 60 if C becomes law. Suppose that all of the interpretations are sufficiently reasonable to withstand *Chevron* if it is applied, but that the interpretations have different probabilities of being approved by a court applying sliding-scale *Skidmore* deference. We can further imagine two hypothetical *Skidmore* regimes, one that is relatively lax and another that is relatively strict, with the probabilities for surviving review being higher for any given interpretation under the former regime than the latter regime. Specifically, imagine that the probabilities of survival in the event *Skidmore* is applied are given by the following table:

	Interpretation A	Interpretation B	Interpretation C
Lax *Skidmore*	20%	65%	95%
Strict *Skidmore*	15%	30%	70%

Finally, suppose that in the *Chevron* lottery, the *ex ante* probability of receiving *Chevron* deference is 60%. Which interpretation does it make sense for the agency to select?

In a world where the *Skidmore* standard applied by courts is lax, the agency's best option is to choose Interpretation B: the middle-of-the-road interpretation that is neither the safest nor the most adventurous.[203] Interpretation B, while not the agency's favorite, is still acceptable to the agency, and it stands a solid chance of being upheld even if *Skidmore* is applied.

On the other hand, if the court applies the stricter version of *Skidmore*, the agency's calculations change: now the best strategy is to choose Interpretation A and hope for *Chevron* deference.[204] The agency must sacrifice so much in policy content to bring its odds of surviving *Skidmore* review above 50% that it makes sense to opt out of the *Skidmore* lottery and wager everything on winning the *Chevron* lottery. Note that this outcome is plainly unsatisfactory: tightening the screws of *Skidmore* has ironically yielded an agency interpretation less faithful to the statutory scheme and also spiked the rate of judicial reversals.[205]

This finding introduces a cautionary note to judicial deference practice. Ratcheting up the scrutiny of *Skidmore* could backfire badly if agencies stand a good chance of drawing *Chevron* deference instead. How much scrutiny is too much will depend on the particulars of agency preferences and the *Chevron* and *Skidmore* lotteries; the following section explores the relations between some of these factors.

2. The Chevron *Lottery and the* Skidmore *Lottery Can Interact to Shape Agency Behavior in Surprising Ways.*—What can be said about how the *Chevron* and *Skidmore* lotteries interact? One might suppose that they are substitutes: that "tightening" the *Chevron* lottery—raising the odds that an agency will face sliding-scale scrutiny—and "tightening" the *Skidmore*

203. The agency's expected benefit from choosing Interpretation B is 68.8: 0.6 * 80 + 0.4 * 0.65 * 80. That is the probability of receiving *Chevron* review (0.6) times the benefit to the agency from Interpretation B (80), plus the probability of receiving *Skidmore* review (0.4) times the probability of surviving that review (0.65) times the benefit to the agency from Interpretation B (80). This compares favorably to the expected benefit from Interpretation A (68) and Interpretation C (58.8).

204. Interpretation A yields an expected benefit of 66. This exceeds the expected benefit from Interpretation B (57.6) or Interpretation C (52.8).

205. We would expect courts to reverse agencies 34% of the time: *Skidmore* will be applied 40% of the time, and the agency will lose 85% of those cases.

lottery—raising the odds an opportunistic interpretation will fail *Skidmore* review—might be equally effective in inducing greater agency compliance.

In fact, it is not possible to make many specific claims about how the *Chevron* and *Skidmore* lotteries interact without knowing more about the agency's utility function and the shape of the *Skidmore* lottery's probability distribution. These are all abstractions, of course, and the fine-grained distinctions that are possible to draw with a mathematical model translate only roughly, at best, to the messier, real-world environment of agencies and courts. Certainly when agencies choose among different possible interpretations, they do not do so by undertaking subjective expected utility calculations with hard numbers.[206] All that being said, it is still worth noting that the *Chevron* and *Skidmore* lotteries interact to shape agency incentives in sometimes counterintuitive ways, and this favors some strategies of judicial review over others.

In particular, it would seem reasonable at first blush to suppose that a strict *Chevron* lottery and a strict *Skidmore* lottery might be effective substitutes in inducing agency compliance. That is, we might expect that a low probability of getting more aggressive review under *Skidmore* could induce the same measure of agency compliance as a higher probability of getting a somewhat lower measure of *Skidmore* scrutiny. In other words, if *Skidmore* is applied quite stringently whenever it is applied, it could be used more sparingly and still induce a desired level of agency compliance.

In fact, the "substitutability" of *Chevron* and *Skidmore* lotteries is not reliable. Making it more likely that agencies will receive *Chevron* does create more agency slack, but tightening up the *Skidmore* scrutiny will not always reduce it. First, recall from above that if *Skidmore* review is too strict, it makes sense for agencies to give up on trying to satisfy it. As a result, ratcheting up *Skidmore* scrutiny to compensate for a looser *Chevron* lottery will backfire if *Skidmore* is pushed past its threshold of effectiveness. Secondly, the interactive effect depends on how adjustments to the intensity of *Skidmore* scrutiny affect the probability of surviving judicial review. Indeed, it is even possible that *decreasing* the intensity of *Skidmore* review may induce more agency compliance.[207]

The broader lesson here is that aggressive review under *Skidmore* is a fairly blunt tool for shaping agency behavior. Ratcheting up the intensity of *Skidmore* review may not reliably rein in agencies, because the incentives generated by the interaction of the *Chevron* and *Skidmore* lotteries vary widely based on the particulars of the situation. On the other hand,

206. *See* Eskridge & Baer, *supra* note 6, at 1091 (concluding that "there is no clear guide as to when the Court will invoke particular deference regimes, and why").

207. This may be the case if the bump upwards in probability of surviving review is greater the closer the agency's interpretation is to the most plausibly faithful interpretation. This kind of manipulation to the *Skidmore* lottery weakens the "stick" (the penalty for opportunistic behavior) but strengthens the "carrot" (the reward for compliant behavior).

tightening up the *Chevron* lottery—that is, making it more likely that agencies will be reviewed under *Skidmore*—will reliably incentivize more compliance from agencies. Taken together, these findings suggest that a moderate intensity *Skidmore*, applied with more frequency, might be a more useful approach to managing agency behavior than very strict *Skidmore* applied sparingly.

This result lines up with arguments, both prescriptive and descriptive, made by other administrative law scholars. Eskridge and Baer call for an overall streamlining of deference doctrine, for smoothing some of the sharp discontinuities between different standards of review in favor of a continuum of deference.[208] A deference lottery that liberally features a moderate, sliding-scale *Skidmore* review, while not something that Eskridge and Baer endorse, enjoys some similarities to their vision.[209] Also, David Zaring has made the descriptive claim that the welter of different judicial review doctrines that courts apply to agencies reduce to a single "reasonableness" standard.[210] This Article does not come to the same conclusion. But to the extent that the deference lottery alternates *Chevron* deference with a moderate intensity *Skidmore* standard applied fairly frequently, the result would be approximately the same.

3. An Unpredictable Chevron *Regime Attenuates* Chevron*'s Capacity to Shape Agency Behavior and Leads to More Judicial Reversals.*—To this point, the analysis has proceeded on the assumption that, if an agency receives *Chevron* review, it is home free: its interpretation will be upheld as a reasonable construction of an ambiguous statute. Of course, in reality, agencies cannot count on surviving *Chevron* review.[211] This section relaxes the assumption and asks how the outcomes change if *Chevron* does not always translate into an agency win. Specifically, this section considers the effect of introducing some random variation into the outcome of *Chevron* review. The consequence is unwelcome: *Chevron*'s power to shape agency behavior goes down, and the rate of judicial reversals goes up. So whereas random assignment to different deference standards can be part of an effective regime for managing agency behavior, random variation in judgment worsens outcomes on any measure.

There is some empirical evidence that once agencies make it past *Chevron* Step One—the question of whether the statute is ambiguous—they are, if not guaranteed to win on Step Two, extremely likely to do so. A study by Orin Kerr finds that, under *Chevron* Step Two, agency interpretations are

208. Eskridge & Baer, *supra* note 6, at 1183–85.

209. There are significant differences as well: for instance, Eskridge and Baer make suggestions for reducing the unpredictability of *Chevron*'s application, and pegging *Skidmore* deference squarely to agency expertise, rather than interpretive content. *See id.* at 1092–93.

210. Zaring, *supra* note 97, at 137.

211. *See, e.g.*, Whitman v. Am. Trucking Ass'ns, 531 U.S. 457, 481–85 (2001) (rejecting agency interpretation at *Chevron* Step Two).

upheld 89% of the time.[212] Imagine, though, that courts applying *Chevron* Step One frequently decide that statutes unambiguously foreclose the agency's interpretation, and that these decisions are not predictable.[213] We can call this situation "Crapshoot *Chevron*." What will be the effect on agency behavior?

The results are straightforward. As the predictability of *Chevron* Step One declines, the *Chevron* lottery's capacity to shape agency behavior attenuates. The more the *Chevron* lottery dissolves into noise, the more a rational agency will key its behavior off the *Skidmore* lottery. Crapshoot *Chevron* can thus have the same effect on agency behavior as increasing the chance of receiving *Skidmore* review—namely, reining in agency interpretations—but with one critical difference: the rate of judicial reversals will rise.[214] To the extent that courts rule against agencies in an unpredictable way, judicial review loses its capacity to guide agency behavior and imposes additional costs in the form of reversals.[215] Any level of agency compliance achieved with Crapshoot *Chevron* could also be achieved under a deference lottery with a fully predictable *Chevron* and at a lower rate of judicial reversals.

Inevitably, there is a certain amount of noise in most doctrinal frameworks, owing to the inherent vagueness of legal standards. *Chevron* analysis, the key operative concepts of which are "ambiguous" and "reasonable," will never be fully determinate or perfectly predictable. But from the perspective of the operation of the deference lottery, it is best to hold the apparent randomness of *Chevron* applications to a bare minimum. A strong default norm for the *Chevron* framework of deciding close questions in the agency's favor might seem to give agencies too much latitude. But in the context of a deference lottery, such a norm supports an effective regime for guiding agency behavior.

---

212. Kerr, *supra* note 193, at 31; *cf.* Hickman & Krueger, *supra* note 125, at 1252 (remarking that "it is unsurprising that most agency interpretations survive *Chevron*'s second step" given that "*Chevron*'s step two nears the fully deferential end of the spectrum").

213. Judges and justices differ in their willingness to grant or refuse deference at *Chevron* Step One. Justice Scalia, for instance, is less inclined to find ambiguity than most of his colleagues. *See, e.g.*, Babbitt v. Sweet Home Chapter of Cmtys. for a Great Or., 515 U.S. 687, 714 (1995) (Scalia, J., dissenting) (arguing that the regulation's interpretation runs afoul of the "unmistakably clear" statute).

214. Hickman & Krueger, *supra* note 125, at 1278 (offering "support for the widely shared belief that *Skidmore* is less deferential than *Chevron*").

215. *See, e.g.*, *Whitman*, 531 U.S. at 486 (reversing in part after rejecting the agency's interpretation).

IV. Conclusion

A. Assessment

This Article has argued that it makes sense to characterize courts' practice of deference to agency statutory interpretations as a lottery with particular features. The Supreme Court's *Mead* ruling offers no clear rule to govern when courts will grant *Chevron* deference. In the event that *Chevron* deference is not forthcoming, *Skidmore* offers the agency no guarantee of survival, but the better the agency can justify its interpretation as consistent with the content of the statute, the better its chances. Part II draws on empirical work to confirm that this characterization of the doctrinal framework is consistent with courts' actual deference practice. That Part first establishes that the most extensive data collected have very little power to predict when a given agency's regulation will receive *Chevron* deference. Second, it establishes that in the circuit courts, *Skidmore* deference is best understood in probabilistic terms, where the agency worsens its odds by choosing constructions that cannot be easily justified with reference to the content of the statute. Part III explores what taking the deference lottery seriously means for how judicial review practices shape agency behavior. The most striking result is that, relative to an all-*Chevron* regime, introducing some chance of getting *Skidmore* review at random can curb agency opportunism—with the important caveat that, if *Skidmore* is too hard to satisfy, it may cease to affect agency behavior altogether, and instead simply result in a higher rate of judicial reversals.

It is not possible, even within the terms of the model described in Part III, to define the optimal configuration of the deference lottery—the ideal mix of *Chevron* and *Skidmore* review, and the ideal level of stringency within *Skidmore*. What is "optimal" depends on what level of agency autonomy in statutory interpretation is the goal, and how costly judicial reversals are thought to be—questions impossible to answer in the abstract. But what the exploration of the workings of the deference lottery suggests is that, if there is some value to curbing agency slack, a deference lottery is not necessarily a bad approach. The Supreme Court's *Mead* decision, which lays down a somewhat vague standard for whether *Chevron* or *Skidmore* applies, has been roundly criticized.[216] This work shows that *Mead*'s vagueness may have hidden virtues. Facing some possibility that they may encounter a standard of scrutiny higher than *Chevron* may induce agencies to take more care in using statutory interpretation to pursue their own goals. When agencies are risk-averse, the effect on their behavior will be stronger still.

This is not to say, however, that if the courts were building a deference regime from scratch, a deference lottery would be the best approach to take. If the goal is to achieve a given level of agency compliance with as few

216. *See supra* notes 75–78 and accompanying text.

judicial reversals as possible, applying a uniform standard of properly calibrated scrutiny would be the best approach. But of course courts are not building a deference regime from scratch; they are working within an existing matrix of precedents. If the baseline norm is that *Chevron* applies whenever agencies interpret statutes in regulations and formal adjudications, as some assumed before *Mead*, the more relevant question is, what effect would it have to introduce the possibility that, sometimes, more scrutiny might be applied? At least in terms of the simple model, the answer is that it may cause agencies to stay closer to the statutory core without causing the reversal rate to spike.

It is also worth noting that the residual, unavoidable unpredictability of judicial standards of review may make the first-best solution—a uniform standard of review, "correctly" calibrated to produce a desired level of agency compliance—a difficult thing to craft. Just to state the goal is to show how elusive its attainment would be. Even if there were agreement in principle as to how much running room agencies should have in construing statutes, what verbal formula would properly express it? And how would it be possible to have a single standard be applied uniformly by the whole appellate bench, particularly given that judges may have preferences over policy and may apply the standard strategically in pursuit of those preferences? A deference lottery acknowledges the irreducible indeterminacy of legal standards and the diversity of the bench, and leverages both of these to produce a regime that can flexibly manage agency behavior. If no single deference formula can reliably find the "sweet spot" of agency autonomy in statutory interpretation, alternation between two different standards, each in the proper proportion, may nudge agencies towards it. The deference lottery is a second-best solution. But we live in the world of the second best, and it may be hard to improve on a deference lottery here.[217]

Two real-world questions naturally arise. The first is, how strict or lax are the deference lotteries being imposed by our courts? The second is, do the lotteries in fact have an effect on the content of agencies' statutory interpretation? I can offer some preliminary thoughts on both questions, although a complete answer to either is well beyond the scope of this Article.

A thorough assessment of the characteristics of the deference lotteries imposed by our courts would require the collection and analysis of new data,

---

217. Another solution that would be effective in principle but difficult to implement in practice—and would also raise troubling questions from a transparency in governance perspective—is to create an "acoustic separation" between how courts review agencies and how agencies *think* courts review agencies. In other words, if the deference regime were in fact quite deferential, but agencies anticipated fairly stringent judicial review, the regime could generate the benefits of agency compliance without the costs of high levels of reversals. *Cf.* Meir Dan-Cohen, *Decision Rules and Conduct Rules: On Acoustic Separation in Criminal Law*, 97 HARV. L. REV. 625, 625, 630 (1984) (defining "acoustic separation" as an imaginary situation in which only officials know the rules for making decisions and only the public knows the conduct rules).

but I can venture some observations based on the data already discussed in this Article. On the whole, it seems that our deference lottery regime pairs a fairly strict *Chevron* lottery with a fairly lax *Skidmore* lottery. In other words, the chance that an agency will get *Skidmore* review is relatively high, and its chances of surviving *Skidmore* review is also relatively high. The Eskridge and Baer data show that, in the Supreme Court at least, agency statutory interpretations in notice-and comment regulations are reviewed either under *Skidmore* or what they term "*Skidmore*-light" approximately 30% of the time.[218] (To the extent that appeals courts do as the Supreme Court says and not as it does, however, the rate at which *Skidmore* is applied could be somewhat lower.)[219] And the Hickman and Krueger work shows that, when the courts of appeals do apply *Skidmore*, they pay careful attention to the justifications offered by agencies, rather than interpreting statutes de novo.[220] The Hickman and Krueger data show that the survival rate for statutory interpretations under *Skidmore* is just over 60%, which suggests that the review is not extraordinarily stringent.[221] From the information available, then, it seems that the deference lottery avoids the combination of extremes—*Skidmore* applied harshly and infrequently—that would cause reversals to mount without reducing agency slack.

But do agencies actually respond to deference lotteries as the theory predicts? This question falls outside the scope of this Article, which is fundamentally a theory-building piece. A thorough empirical analysis would require either in-depth case studies of agency decision making or a large-scale quantitative analysis of the content of agency statutory interpretations, both of which present formidable challenges of data collection and measurement.

That being said, the agency behaviors I posit here are plausible, based on what we already know about how agencies operate. We know that agencies' leaders do care how their actions fare in courts, and that agencies make choices with an eye to surviving judicial review. Indeed, the prominent scholarship from the 1980s and 1990s on the "ossification" of rulemaking demonstrates that agencies respond strategically to cues from the judiciary. That body of work demonstrates in detail how intensive judicial

---

218. *See supra* text accompanying note 104.

219. As discussed above, the language of *Mead* suggests a strong presumption that *Chevron* will apply to statutory interpretations announced in notice-and-comment rulemakings and formal adjudications. *See supra* notes 70–71.

220. *See* Hickman & Krueger, *supra* note 125, at 1247 (stating that the *Skidmore* deference standard requires "reviewing courts to evaluate an interpretation's persuasiveness by weighing various factors").

221. *Id.* at 1275. Note, however, that the population of cases in the Hickman and Krueger dataset includes *all* statutory interpretations, not only those promulgated through notice-and-comment rulemaking or formal adjudication. Since the default standard for agency statutory interpretations promulgated through informal means is *Skidmore*, it may be that agencies are more conservative with these interpretations, pushing the survival rate up.

scrutiny of rulemaking[222] can cause agencies either to forego rulemaking in favor of other forms of activity,[223] or else to invest in additional procedures[224] or explanation[225] in order to pad out the record for judicial review. Moreover, we know that the General Counsel's Office, which presumably keeps abreast of developments in judicial review of agencies, is involved in major policy initiatives from the earliest stages, at least in some large agencies.[226] For the model to reflect actual agency practice, all that needs to happen is that personnel within agencies are broadly aware of reviewing courts' recent deference practices with respect to the agency—do they grant *Chevron* review frequently, and if not, how stringent does review tend to be?—and that they bring this knowledge to bear when policy is made. And it seems that changes in judicial deference practices do, in fact, induce changes in how agencies interpret statutes, at least some of the time. Donald Elliott, a former General Counsel for EPA, reports that "[EPA] and other agencies gradually internalized and adapted to the additional interpretive discretion (i.e., the expanded power) that *Chevron* provided them. Accordingly, EPA and other agencies are now more adventurous when interpreting and elaborating statutory law."[227]

---

222. Richard J. Pierce, Jr., *Two Problems in Administrative Law: Political Polarity on the District of Columbia Circuit and Judicial Deterrence of Rulemaking*, 1988 DUKE L.J. 300, 300–01 (remarking that "policymaking through agency rulemaking has declined significantly at some agencies during the past decade," in large part because of "the approach taken by appellate courts when they review agency rules").

223. *See* MASHAW & HARFST, *supra* note 171, at 95, 148–49 (describing how the National Highway Transportation Safety Administration abandoned rulemaking in favor of vehicle recalls as a tool for enhancing public safety, in large part due to the inhospitable reception of its rules by the circuit courts).

224. *See* Thomas O. McGarity, *Some Thoughts on "Deossifying" the Rulemaking Process*, 41 DUKE L.J. 1385, 1410–26 (1992) (describing how agencies hold additional hearings, convene panels of outside experts, and undertake studies to anticipate judicial challenges to rulemakings).

225. *See* Mashaw, *supra* note 14, at 196, 197 n.38 (noting the rapid growth, between the 1970s and 1990s, in the length of the "concise statement[s] of basis and purpose" that the Administrative Procedure Act requires agencies to file in connection with rulemakings).

226. *See* Magill & Vermeule, *supra* note 185, at 1079–80 (observing how doctrines that extend the scope of judicial review increase the leverage of lawyers over agency policy-making processes); *see also* Thomas O. McGarity, *The Internal Structure of EPA Rulemaking*, LAW & CONTEMP. PROBS., Autumn 1991, at 57, 63–90 (providing an extensive overview of the EPA's internal decision-making process, including the role of agency lawyers); *id.* at 67 ("[Office of General Counsel's] duty to ensure that rules survive 'arbitrary and capricious' review justifies the office in taking positions on the substantive merits of proposals and on the technical and economic validity of the support documents.").

227. E. Donald Elliott, Chevron *Matters: How the* Chevron *Doctrine Redefined the Roles of Congress, Courts and Agencies in Environmental Law*, 16 VILL. ENVTL. L.J. 1, 3 (2005) (footnotes omitted). That said, the impact of changes in judicial doctrine on agency behavior should not be overstated. Several years earlier, Elliott himself was quoted saying, "I would take issue with the assertion that we know that the effects of judicial review on the administrative process and on the internal deliberations within agencies are huge." *Administrative Law Symposium: Question & Answer with Professors Elliott, Strauss, and Sunstein*, 1989 DUKE L.J. 551, 553.

B. Recommendations

Still, it is possible to offer some recommendations for some tweaks to courts' practice that can improve the performance of the deference lottery, whatever level of latitude is ideal for agencies to have. When making recommendations, it is important to bear in mind that no single individual or entity is in charge of defining the contours of the deference lottery; rather, it is a phenomenon that emerges from the interactions of independent decisions made by multiple judges, and indeed depends on judges having different patterns of behavior. All that being said, the Supreme Court plays a critical role in setting out the argumentation frameworks that shape how all federal courts tackle legal questions.[228] And there are two subtle changes to the Supreme Court's deference doctrines that would make the deference doctrine more effective at directing agency behavior, by bringing actual practice more in line with the assumptions made in my model.

The first would be to reinforce the "sliding-scale" variant of *Skidmore* analysis, which best encourages agencies to strive for interpretive fidelity to the statutes they administer. As discussed above,[229] close study of the appellate courts has identified two major strains of *Skidmore* analysis, the sliding-scale model and the independent-judgment model. The sliding-scale model better rewards agents for more justifiable readings of the statutes they administer. Under sliding-scale review, the agency's interpretation, rather than the text of the statute, is the starting point for the court's analysis, and it will stand or fall depending on how convincing a case the agency can make for it. Even if an interpretation is not the one the court would have chosen *ab initio*, the court is open to the agency's reasons for its choice and will credit those reasons proportional to their power to persuade. This form of analysis trains the reviewing court's focus squarely on the relevant question from an agency theory perspective: not, what interpretation would the court choose, but how justifiable is the interpretation chosen by the agency?

Although the factors expressly named in *Skidmore* do not speak directly to agency expertise,[230] some commentators have understood *Skidmore* to peg deference to expertise,[231] and courts have sometimes applied it that way as well.[232] There may be good reasons to defer more to agencies with strong subject-specific expertise, but focusing exclusively on expertise leaves agencies no incentive to subvert their own policy preferences in favor of fidelity to the statutes they administer. Even if courts consider expertise in

---

228. On argumentation frameworks in law, see generally Alec Stone Sweet, *Path Dependence, Precedent, and Judicial Power*, *in* ON LAW, POLITICS, & JUDICIALIZATION 112 (Martin Shapiro & Alec Stone Sweet eds., 2002) and Giovanni Sartor, *A Formal Model of Legal Argumentation*, 7 RATIO JURIS 177 (1994).

229. *See supra* section II(B)(2).

230. Skidmore v. Swift & Co., 323 U.S. 134, 140 (1944).

231. Krotoszynski, *supra* note 137, at 754.

232. *See* Hickman & Krueger, *supra* note 125, at 1288–89 (noting that many courts' deferential application of the *Skidmore* standard considered agency expertise).

their *Skidmore* analysis, courts should also consider the content of agencies' interpretations and the justifications agencies offer for them. As Hickman and Krueger found, the sliding-scale model of *Skidmore* review is already dominant on the circuit courts,[233] although both the circuit courts and the Supreme Court rely on their independent judgment in a significant share of cases.[234] To the extent that the Supreme Court can signal in its decisions that this is at the core of *Skidmore* analysis, the *Skidmore* lottery can exert a more consistent and effective pull on agency behavior.

Also, as I discussed above, the deference lottery works best if, when *Chevron* is applied, it is applied predictably and with a good deal of deference. There are three ways in which courts could diverge from this ideal. On the one hand, rather than deferring to any reasonable interpretation, broadly construed, courts could apply more scrutiny to the agency's interpretation and rationale, so that *Chevron*, too, functions as a kind of sliding-scale review. Convergence between *Chevron* and *Skidmore* is not necessarily undesirable,[235] but it may induce agencies to play it safer with their interpretations than is optimal. When *Chevron* becomes *Skidmore*-light, the deference lottery becomes less flexible as a tool to regulate agency behavior. The second divergence from the idealized *Chevron* outlined here is the "Crapshoot *Chevron*" described in subpart III(C), in which the outcomes of judicial review within *Chevron* vary unpredictably. Crapshoot *Chevron* likewise reduces the deference lottery's ability to shape agency behavior, and it also translates into more judicial reversals with no gains in terms of agency fidelity.[236] The third divergence is that courts can give *Chevron* deference at different rates to different agencies. We know that there is currently cross-agency variation within the *Chevron* lottery.[237] To the extent that agencies face different *Chevron* lotteries, of course, the deference regime as a whole gives them different incentives. Barring grounds for differential treatment—such as a judicial judgment that some agencies can be trusted with discretion more than others—the Supreme Court would do well to subject all agencies to the same lottery, and there are some signs that such a convergence is underway.[238]

In a legal regime, clarity has value, but sometimes unpredictability does too. This Article has argued that the deference regime agencies face when

233. *Id.* at 1238.

234. *Id.*; *see also* Eskridge & Baer, *supra* note 6, at 1090 ("[I]n the majority of cases—53.6% of them—the Court does not apply any deference regime at all."). When we confine our attention to agency statutory interpretations offered in formal adjudications or notice-and-comment rulemakings, the Eskridge and Baer data show that the Supreme Court still reviews the agency without reference to any deference standard in 16% of the cases.

235. *See* Zaring, *supra* note 97, at 137 (arguing that the various standards for reviewing agency action have already converged into a single "reasonableness" standard).

236. *See supra* notes 214–30 and accompanying text.

237. *See supra* note 107 and accompanying text.

238. *See supra* note 107.

they seek to defend their statutory interpretations in court amounts to a lottery, and that a lottery can be an effective tool for managing agency behavior. If it were possible to craft legal standards with laser-like precision, if there were no variability in how judges applied standards, and if courts were devising a deference regime against a blank slate, there would be little to recommend a deference lottery. But given the incomplete determinacy of any legal standard, the variability in judicial behavior, and *Chevron*'s place in precedent as a default rule, the deference lottery approach may be the best available option for appropriately structuring the relationship between courts and agencies.

Book Reviews

Copyright's Cultural Turn

CONFIGURING THE NETWORKED SELF: LAW, CODE, AND THE PLAY OF EVERYDAY PRACTICE. By Julie E. Cohen. New Haven, Connecticut: Yale University Press, 2012. 352 pages. $55.00.

Reviewed by Anupam Chander* & Madhavi Sunder**

Introduction

How ironic that the scholarship on the area of law most directly regulating the culture industries has long resisted learning from scholarship on culture! Rather than turning to cultural studies, anthropology, geography, literary theory, science and technology studies, and media studies, over the last few decades copyright scholars have relied largely on economics for methodology.

However, the hegemony of law and economics in copyright is yielding. The exhortation of some of this school to commodify creativity to render it market tradable is increasingly exposed as deficient both as a sufficient mechanism to improve people's lives and as a vision of what makes a life good in the first place. Most importantly, by failing to recognize the importance of creative works beyond their economic value, a policy dictated

* Director of the California International Law Center and Professor of Law, University of California, Davis; J.D. Yale Law School, A.B. Harvard College.

** Professor of Law, University of California, Davis; J.D. Stanford Law School, A.B. Harvard College. For insightful conversations, we thank Shyam Balganesh, Mario Biagioli, Margaret Chon, Haochen Sun, and Talha Syed. We are grateful to Carl Larson and Christine Meeuwsen for very helpful research assistance.

by economic analysis alone might fail to provide sufficient limits on the rights of copyright holders.

Julie Cohen's new book, *Configuring the Networked Self: Law, Code, and the Play of Everyday Practice*,[1] marks a major effort to craft a jurisprudence of information law that goes beyond law and economics. Cohen, a celebrated scholar of intellectual property and privacy, brings her formidable talents to the fore in this book to ask scholars in both fields to pay more attention to culture. Cohen argues that the dominant approach to copyright and privacy fails to understand the role of information in people's actual lives. We have become too enamored with abstract claims of human behavior that turn out to be incomplete upon closer examination, she tells us. Mining a broad vein of contemporary theory ranging from science and technology studies to cultural studies, Cohen seeks to inform policy on intellectual property and privacy with an understanding of what she calls the networked self, the individual embedded in a complex structure of social and technological circumstances.[2]

Cohen's book is part of what we believe to be a "cultural turn" in intellectual property thinking. Her book is part of an emerging school of analysis, which brings interdisciplinary insights from fields other than economics to explore the deeper significance and role of cultural products. Beginning with Rosemary Coombe and Keith Aoki, legal scholars have sought to learn from the humanities and social sciences beyond economics to better understand why we create, how we create, who creates, and the effects of cultural production on social and economic well-being.[3] Increasingly, scholars writing in this vein draw their normative vision from the work of Martha Nussbaum and Amartya Sen, who drew attention to the need to improve quality of life by enhancing the capabilities of each person.[4] An intellectual property policy would thus be evaluated by a new metric, not

---

1. JULIE E. COHEN, CONFIGURING THE NETWORKED SELF: LAW, CODE, AND THE PLAY OF EVERYDAY PRACTICE (2012).

2. *Id.* at 6–8.

3. *See generally* ROSEMARY J. COOMBE, THE CULTURAL LIFE OF INTELLECTUAL PROPERTIES: AUTHORSHIP, APPROPRIATION, AND THE LAW (1998); Keith Aoki, *(Intellectual) Property and Sovereignty: Notes Toward a Cultural Geography of Authorship*, 48 STAN. L. REV. 1293, 1355 (1996) (calling for recognition of "hybridities, pluralisms, and localisms" when considering intellectual property law).

4. *See, e.g.*, MADHAVI SUNDER, FROM GOODS TO A GOOD LIFE: INTELLECTUAL PROPERTY AND GLOBAL JUSTICE 7 (2012) [hereinafter SUNDER, FROM GOODS TO A GOOD LIFE] (drawing upon the work of Sen and Nussbaum to consider how intellectual property laws can promote human freedom and development); Margaret Chon, *Intellectual Property and the Development Divide*, 27 CARDOZO L. REV. 2821, 2823 (2006) (proposing a substantive equality principle to guide global intellectual property policy making); Brett Frischmann & Mark P. McKenna, *Intergenerational Progress*, 2011 WIS. L. REV. 123, 137; Lea Bishop Shaver, *Defining and Measuring A2K: A Blueprint for an Index of Access to Knowledge*, 4 I/S: J.L. & POL'Y FOR THE INFO. SOC'Y 235, 239 (2008); Madhavi Sunder, *IP*^{3}, 59 STAN. L. REV. 257, 313–15 (2006) [hereinafter Sunder, *IP*^{3}] (applying the capabilities approach to conflicts in intellectual property law).

simply increased products (in the form of patents, copyrighted works, or trademarked goods), or its contribution to the gross domestic product, but rather its role in enhancing human capabilities. Rejecting the stylized utilitarianism of law and economics,[5] Cohen explicitly embraces the capabilities approach of Nussbaum and Sen.[6]

Cohen's book defies easy summary, and we do not seek to do so here. It covers a broad legal landscape from copyright and privacy to communications policy. She critiques liberal policies for what she sees as their inattention to the endogeneity of the individual self, that is, the dialectic process between culture and subjectivity, with each influencing the other.[7] She argues for the importance of play as a "vital catalyst of creative practice, subject formation, and material and spatial practice."[8] Cohen seems to define "play" as not rigid, rather than not work.[9] One of her primary concerns is the inevitable creep toward total control (legal, cultural, and technological) of a digital information society. Cohen advocates, instead, for flexibility and gaps in the digital networked environment because these interstitial spaces are where creativity and self-formation may fruitfully occur.[10] She offers three strategies to enhance the possibility of play: access to knowledge, operational transparency, and semantic discontinuity.[11] The first two strategies are largely well-known, but the third requires elaboration. By semantic discontinuity, Cohen means an incompleteness in the legal and technical landscape that leaves unregulated spaces for individual action.[12]

In an early review, Jack Balkin agrees with Cohen that we all need what he calls "room for maneuver," but worries that semantic discontinuity may be insufficient to offer this space without more planned policy making.[13]

---

5. COHEN, *supra* note 1, at 21 ("An adequate theoretical framework for information law and policy must allow the definition of rights without insisting that they be amenable to neutral, quasi-scientific reduction, and must permit formulation and discussion of instrumental goals without imposing the Procrustean requirements of utilitarianism.").

6. *Id.* at 21 ("The theory of capabilities for human flourishing satisfies both requirements, and supplies the underlying normative orientation for the analysis developed in this book.").

7. *Id.* at 7.

8. *Id.* at 223.

9. *Id.* at 55.

10. *Id.* at 227.

11. *Id.* at 31.

12. *See id.* (defining "semantic discontinuity" as "an interstitial complexity that prevents the imposition of a highly articulated grid of rationality on human behavior and instead creates spaces within which the play of everyday practice can move").

13. As Jack Balkin has observed:

> First, semantic discontinuity might be only a second-best solution to the problem of freedom. Surely one would want at least some rules, technologies, and practices that directly protected individuals from overreaching by powerful public and private entities. . . .
>
> . . . .

Anita Allen has suggested that Cohen's views might not be as hostile to liberalism as she suggests.[14]

We seek here to flesh out Cohen's important arguments in two ways—first, by contextualizing them through comparison with the reigning law and economics approach; and second, by highlighting some key insights of a cultural analysis of copyright. (We confine our arguments to intellectual property, not taking up Cohen's ambitious undertaking to analyze privacy under the same umbrella.) Cohen herself does not frame her approach as a contrast to law and economics. But given the dominance of that approach in legal scholarship, Cohen's book marks a major methodological departure. Cohen writes, "The mainstream of debate about copyright theory and policy . . . tends to ignore or discount the well-established humanities and social science methodologies that are available for investigating the origins of artistic and cultural innovation."[15] In addition to embracing the normative goals of enhancing play and realizing the networked self, the major contribution of her book is to broaden the methodological tools available for analyzing intellectual property policy.

Here, we further develop a cultural approach to intellectual property policy that focuses on expanding human capabilities. Our goal is not to replace law and economics with another, allegedly complete jurisprudential system, but to supplement it with a broader set of disciplines with which to understand our world and to allow greater questioning of the ideological entailments of any particular jurisprudential approach. The capabilities approach does not repudiate economics, but simply changes the metrics for judging economic progress and development. Sen, after all, earned his Nobel prize in economics.[16]

> . . . . Gaps and ambiguities in code and law that benefit individuals might also benefit powerful corporations, and vice versa.

Jack M. Balkin, *Room for Maneuver: Julie Cohen's Theory of Freedom in the Information State*, 6 JERUSALEM REV. LEGAL STUD. 84–85 (2012).

14. Anita Allen, *Configuring the Networked Self: Shared Conceptions and Critiques*, CONCURRING OPINIONS (Mar. 6, 2012, 6:14 PM), http://www.concurringopinions.com/archives/2012/03/configuring-the-networked-self-shared-conceptions-and-critiques.html#more-59028; *see also* MARTHA C. NUSSBAUM, CREATING CAPABILITIES: THE HUMAN DEVELOPMENT APPROACH 35 (2011) ("Capabilities belong first and foremost to individual persons, and only derivatively to groups."); Amartya Sen, *The Impossibility of a Paretian Liberal*, 78 J. POL. ECON. 152, 152–53 (1970) (examining the consequences associated with the concept of individual liberty). Cohen embraces Nussbaum's normative vision, but Nussbaum herself is avowedly liberal, as is Sen. The difference may be in how each characterizes liberalism. Nussbaum and Sen see it as an approach that embraces individual definition of what constitutes a good life, while Cohen worries that liberalism relies upon the mistaken view that individuals are autonomous beings, capable of such self-definition. The divergence between the views may not prove practically decisive. Cohen's policy prescriptions seem to largely track traditional liberal ones.

15. COHEN, *supra* note 1, at 18.

16. *The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 1998*, NOBEL PRIZE, http://www.nobelprize.org/nobel_prizes/economics/laureates/1998/index.html.

We follow Cohen's call to situate intellectual property policy in the lives of real people,[17] by imagining how intellectual property might affect the life of Vasanti, "a small woman in her early thirties who lives in Ahmedabad."[18] Martha Nussbaum introduces Vasanti in writing on how to create capabilities.[19] Vasanti is illiterate, without resources of her own, having left her abusive husband, and earns a meager income from "making eyeholes for the hooks on sari tops."[20] As Nussbaum describes, Vasanti's life chances improved dramatically with a loan from the Self-Employed Women's Association, a world-class not-for-profit organization that happened to be based in Vasanti's hometown.[21] As Nussbaum shows through her focus on Vasanti, the capabilities approach is inherently focused on people's actual lives. We note that our discussion of Vasanti is hypothetical, lacking the ethnographic realism of Nussbaum's work, science and technology studies, or Cohen's ideal approach.

Part I reviews some of the principal deficiencies of law and economics as a complete method for intellectual property policy making. Part II seeks to go beyond economics by articulating how a cultural approach focused on enhancing human capabilities would change the ways we understand and regulate cultural production and exchange.

I. Why Economics Is Not Enough

The two principal deficiencies of the law and economics approach are both well-known. First, the foundational understanding that monopoly rights on information are generally necessary to induce the creation of that information has been called into question by seemingly innumerable sources. Second, a single-minded focus on efficiency neglects the distribution of resources in society.

A. Why Do Writers Write?

What justifies copyright law? For scholars writing from the perspective of law and economics, we need copyright law because of market failures that would prevail in its absence.[22] Without copyrights, authors would not write because their creations would simply be copied freely by others without any

17. COHEN, *supra* note 1, at 4–6.

18. NUSSBAUM, *supra* note 14, at 2; *see also* MARTHA C. NUSSBAUM, WOMEN AND HUMAN DEVELOPMENT: THE CAPABILITIES APPROACH 16 (2000) [hereinafter NUSSBAUM, WOMEN AND HUMAN DEVELOPMENT].

19. NUSSBAUM, *supra* note 14, at 2–6.

20. *Id.* at 2. Vasanti's occupation seems to epitomize the division of labor described by Adam Smith a century and a half earlier.

21. *Id.*

22. *See* Wendy J. Gordon, *An Inquiry into the Merits of Copyright: The Challenges of Constistency, Consent, and Encouragement Theory*, 41 STAN. L. REV. 1343, 1435 (1989) (describing the market failures that would ensue in a world without clearly defined property rights).

monetary benefit to the authors.[23] Lacking remuneration available through enforceable rights, creativity would grind to a halt.[24] As Cohen writes, "[B]oth copyright lawyers and copyright scholars tend to assume that copyright law is centrally important in stimulating a high level of creativity."[25] While it is reasonable to argue that the millions of dollars required to develop a new software package, video game, or movie might not be forthcoming were it not for the promise of a monetary reward protected by a copyright, it is not so clear that music and books would not be written without this inducement.

Scholars have questioned the claim that creativity falters without monetary reward. Yochai Benkler has observed that direct monetary incentives proved unnecessary for the creation of enormous software packages such as Linux or knowledge resources such as Wikipedia.[26] Eric Von Hippel, Kal Raustiala, and Chris Sprigman have shown how a variety of industries exhibit creativity in the absence of effective copyright protections.[27] Reviewing psychological studies of creativity, Diane Zimmerman, Jeanne Fromer, and Greg Mandel show that economic incentives are often not the driving force behind creativity.[28] Rebecca Tushnet shows that

---

23. *See id.* at 1435–36 (outlining the free rider problem).

24. Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 558 (1985) ("By establishing a marketable right to the use of one's expression, copyright supplies the economic incentive to create and disseminate ideas."); WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW 40 (2003) ("In the absence of copyright protection the market price of a book or other expressive work will eventually be bid down to the marginal cost of copying, with the result that the work may not be produced in the first place because the author and publisher may not be able to recover their costs of creating it."); ROBERT P. MERGES ET AL., INTELLECTUAL PROPERTY IN THE NEW TECHNOLOGICAL AGE 14 (4th ed. 2006) ("Intellectual property protection is necessary to encourage inventors, authors, and artists to invest in the process of creation. Without such protection, others could copy or otherwise imitate the intellectual work without incurring the costs and effort of creation, thereby inhibiting the original creators from reaping a reasonable return on their investment."); Gordon, *supra* note 22, at 1348 ("That economics should be a focus of attention is unsurprising, since both copyright and patent law are seen as serving primarily economic incentive functions."); Mark A. Lemley, *Private Property*, 52 STAN. L. REV. 1545, 1550 (2000) ("By and large, intellectual property exists only where there is a public goods problem—where people need incentives to invest in the creation of new things."); Maureen A. O'Rourke, *Drawing the Boundary Between Copyright and Contract: Copyright Preemption of Software License Terms*, 45 DUKE L.J. 479, 484 (1995) ("Traditional literary works such as books resemble public goods in that an author is unlikely to make the investment to create the book if all may copy it without fee upon its publication.").

25. COHEN, *supra* note 1, at 100.

26. YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM 5–6 (2006).

27. *See generally* Kal Raustiala & Christopher Sprigman, *The Piracy Paradox: Innovation and Intellectual Property in Fashion Design*, 92 VA. L. REV. 1687 (2006); Emmanuelle Fauchart & Eric von Hippel, *Norms-Based Intellectual Property Systems: The Case of French Chefs* (MIT Sloan Sch. of Mgmt., Working Paper No. 4576-06, 2006).

28. Jeanne C. Fromer, *A Psychology of Intellectual Property*, 104 N.W. U. L. REV. 1441, 1443–44 (2010); Gregory N. Mandel, *Left-Brain Versus Right-Brain: Competing Conceptions of Creativity in Intellectual Property Law*, 44 U.C. DAVIS L. REV. 283, 285–86 (2010); Diane

"artists' own experiences of creation" often reveal a desire to create rather than an economic motivation.[29] Eric Johnson concludes that "the social science literature leads to the identification of a general rule that intellectual labors will tend to flourish naturally, without external rewards."[30] None of this suggests that money is irrelevant, only that focusing on it exclusively neglects a significant amount of human creation and motivation.

Not only have scholars undermined the incentive theory's empirical foundation, they have also pointed out the costs of a single-minded focus on propertization. James Boyle and Carol Rose have observed the central role of the public domain of information in enriching our lives, a role often forgotten in the headlong rush to commodify.[31] Jessica Litman, Peter Jaszi, Keith Aoki, David Lange, Peter Lee, and Brett Frischmann have demonstrated the importance of the public domain to downstream innovation and creativity, renewing the adage that we all stand on the shoulders of giants.[32] The recognition of the essential importance of the public domain counters the call for increasing commodification; this scholarship counters the view that if the public domain had been properly parceled out, it would have been deployed in the most efficient manner.

B. Who Gets What?

Martha Nussbaum tells a story from Charles Dickens' *Hard Times* to illustrate a central failing of utilitarianism.[33] Circus girl Sissy Jupe is asked by her teacher to imagine herself in a nation where there are "fifty millions of money."[34] The teacher inquires, "[I]sn't this a prosperous nation, and a'n't you in a thriving state?"[35] Sissy does not know how to answer the question.[36]

Leenheer Zimmerman, *Copyrights as Incentives: Did We Just Imagine That?*, 12 THEORETICAL INQUIRIES L. 29, 29 (2011).

29. Rebecca Tushnet, *Economies of Desire: Fair Use and Marketplace Assumptions*, 51 WM. & MARY L. REV. 513, 515 (2009).

30. Eric E. Johnson, *Intellectual Property and the Incentive Fallacy*, 39 FLA. ST. U. L. REV. 623, 627 (2012).

31. James Boyle, *The Second Enclosure Movement and the Construction of the Public Domain*, LAW & CONTEMP. PROBS., Winter/Spring 2003, at 33, 38–39; Carol M. Rose, *Romans, Roads, and Romantic Creators: Traditions of Public Property in the Information Age*, LAW & CONTEMP. PROBS., Winter/Spring 2003, at 89, 89–90.

32. *See* Keith Aoki, *Authors, Inventors and Trademark Owners: Private Intellectual Property and the Public Domain (Part I)*, 18 COLUM.-VLA J.L. & ARTS 1, 2–3 (1994) (advocating for fewer copyright restrictions on artists); David Lange, *Recognizing the Public Domain*, LAW & CONTEMP. PROBS., Autumn 1981, at 147, 176 (arguing against privatizing the public domain); Peter Lee, *Toward a Distributive Commons in Patent Law*, 2009 WIS. L. REV. 917, 917 (contending that the public domain can more effectively increase access to downstream patented health technologies for low-income communities); Jessica Litman, *The Public Domain*, 39 EMORY L.J. 965, 969 (1990) (asserting that the public domain can solve problems of authorship).

33. NUSSBAUM, WOMEN AND HUMAN DEVELOPMENT, *supra* note 18, at 60.

34. CHARLES DICKENS, HARD TIMES 42 (Paul Negri & Kathy Casey eds., 2001).

35. *Id.*

She later tearfully explains to a friend that she could not answer the question "unless I knew who had got the money, and whether any of it was mine."[37]

The traditional economic approach to intellectual property fails to pay attention to the just distribution of the benefits of intellectual property. This distributional inattention leads to a number of results. First, enthralled with the single motivation of granting strong property rights to authors to induce creation, such approaches fail to value the contributions and concerns of potential users. Second, a singular focus on ability and willingness to pay will induce the creation of the works sought by those with some degree of market power—leading, for example, to the production of many drugs to treat baldness but few remedies for malaria. The inventions and works most useful to the poorest are forgotten under this theory, often lacking sufficient market incentive to induce their creation. Finally, the poorest may lack the ability to access creative works that are protected by globalized exclusionary laws—laws that fence them out.

While utilitarianism can build in some distributional concerns through such features as the diminishing marginal utility of the dollar,[38] the wealth-oriented approach championed by William Landes and Richard Posner lacks even that feature.[39]

II. What Do Cultural Studies Teach Us?

If not economics alone, then what else? Other disciplines in social science and humanities can supplement our effort to understand the role of intellectual property in the lives of people like Vasanti. The introduction of psychology, sociology, cultural studies, literary theory, geography, anthropology, performance, and science and technology studies does not render economics irrelevant. We seek not to *supplant* economics, but to *supplement* it from insights in other academic studies. Indeed, Julie Cohen canvasses scholarship in all of these fields in order to better understand the reality of people's everyday lives. Economics alone among academic fields cannot supply the insights needed to define information policy. In addition to paying attention to supply and demand curves and deadweight loss, we

---

36. *Id.*

37. *Id.* at 42–43.

38. William W. Fisher & Talha Syed, *Global Justice in Healthcare: Developing Drugs for the Developing World*, 40 U.C. DAVIS L. REV. 581, 603 (2007) (proposing that utilitarianism can be egalitarian and explaining, "[W]hen combined with weak and plausible assumptions of diminishing marginal utility and randomized distribution of utility functions, it tends toward a rough egalitarianism, at least with respect to the distribution of basic resources or goods.").

39. Matthew D. Adler, *Cost-Benefit Analysis, Static Efficiency, and the Goals of Environmental Law*, 31 B.C. ENVTL. AFF. L. REV. 591, 593 (2004) ("[C]onsider that a transfer of wealth from rich to poor is not going to be Kaldor-Hicks efficient, or pass a cost-benefit test traditionally understood, but it will increase overall well-being assuming that—as seems quite plausible—money has diminishing marginal utility.").

need to develop and analyze ethnographies, quantitative and qualitative empirical research, psychologies, and sociologies of intellectual property.

Cohen's work is part of what we might term the *cultural* turn in intellectual property law. We identify here two central insights of the cultural turn in intellectual property scholarship: the relationship between cultural products and the self, and the relationship between culture and human development, which we might characterize as the relationship between goods and a good life.[40]

Neither the stylized model of human behavior nor distributional neglect marks the most significant deficiency of the economics approach to copyright. As one of us (Sunder) has written, "The fundamental failure in the economic story of intellectual property has to do with information's role in cultural life and human flourishing."[41] Cohen's book, like the work of Rosemary Coombe before her,[42] seeks to better understand the way that creative works affect us. Understanding the cultural life of intellectual property (to borrow Coombe's wonderful phrase) helps us recognize that creative works are not just passively consumed objects unrelated to human subjectivity. Cultural works are raw materials from which we form ourselves and societies.

The traditional law and economics approach to copyright imagines a stylized world in which the end goal is to satisfy individual preferences by creating works those individuals desire. Understood in this way, the goal of copyright law thus becomes the creation of products for our consumption. This neglects the interplay of the cultural works with people and with each other. But what if we understood creative works as crucial to education, socialization, and even the creation of our own identities?[43]

A. The Situated, Networked Self

Cohen's account is particularly helpful in elaborating the latter connection. "[C]ulture is not a fixed collection of texts and practices," she writes, "but rather an emergent, historically and materially contingent process through which understandings of self and society are formed and re-formed."[44] To use the popular terminology of Bruno Latour, human beings

40. *See* SUNDER, FROM GOODS TO A GOOD LIFE, *supra* note 4, at 31–44 (criticizing the tendency of intellectual property scholars to focus only on the proper alignment of economic incentives and introducing a cultural intellectual property framework).

41. *Id.* at 31.

42. COOMBE, *supra* note 3.

43. *See* SUNDER, FROM GOODS TO A GOOD LIFE, *supra* note 4, at 64–76 ("Participatory culture is instrumentally and intrinsically related to promoting freedom, engendering equality, and fostering human and economic development.").

44. COHEN, *supra* note 1, at 25.

are "hybrids" of the techno-cultural milieus in which we live.[45] Culture and technology shape us as much as we shape them.

One of the central insights of the new cultural studies of intellectual property centers on the relationship between goods and persons. Cohen's view of "culture" takes seriously the constitutive role of technologies and cultural artifacts in configuring the self. "Our beliefs, goals, and capabilities are shaped by the cultural products that we encounter, the tools that we use, and the framing expectations of social institutions," Cohen writes.[46] Cohen is highly influenced by Science and Technology Studies (STS), which posits selves as hybrids of technology, goods, and ideologies.[47] At the same time, as Cohen argues, selves are not passive receptors of technologies, but are dynamic agents in a back and forth with technologies.[48]

The situated, networked self stands in contrast to the liberal self who makes her life in opposition to or outside the boundaries of culture. The situated self is an endogenous creation of the system itself. Selves and technologies are mutually engaged in recursive processes of creation and recreation.

Cohen moves from describing the imbrication of self formation and culture to offering some thoughts on how and why law ought to direct this relationship. She is first and foremost concerned with the freedom-enhancing function of what she calls the "play of everyday practice."[49] Individuals, she argues, ought not be too constricted in their technological and cultural play.[50] She calls this flexibility semantic discontinuity.[51] Notably, Cohen's calls for semantic discontinuity or more room for play are not motivated by a singular desire to promote more innovation or creative expression.[52] She views play in cultural worlds as essential to personal

45. *See* BRUNO LATOUR, WE HAVE NEVER BEEN MODERN 3 (Catherine Porter trans., 1991) (describing hybrids as "half engineers and half philosophers" who attempt to navigate the interconnectivity of science and culture). *See generally* Alain Pottage, *The Materiality of What?*, 39 J.L. & SOC'Y 167 (2012) (elaborating on the insights of actor-network theories in Science and Technology Studies, particularly Latour's theory).

46. COHEN, *supra* note 1, at 2.

47. *See id.* at 25 ("The approaches that I identify as most pertinent . . . focus careful, critical attention on the 'hybrid' assemblages that emerge where politics, economics, technology, ideology, and discourse intersect.").

48. *See id.* at 50 ("Embodied, situated users interact with networked information technologies on a day-to-day basis, often turning those technologies to new purposes and adapting them in unexpected ways.").

49. *See id.* at 50–57 (stressing the importance of understanding the "ordinary, everyday ways that people use information").

50. *See id.* at 57 ("[T]he play of everyday practice is the means by which human beings flourish. . . . It therefore must be a central consideration in evaluating the constellations of legal, institutional, and technical developments with which this book is concerned.").

51. *Id.* at 239–41.

52. *Id.* at 227.

freedom and *self*-creation.[53] She notes, the "reservation of authority to shape the material conditions of everyday life promotes both innovation and psychological and social well-being."[54] For Cohen, the more interesting benefits of cultural play are unexpected; she prizes either play for play's sake or accidental innovation that arises from cultural play.

Even this description of the importance of cultural play may not go far enough. Freedom in the cultural sphere is as important, if not more so, as freedom in the political sphere. The fact that cultural images and values are so powerful a factor in shaping selves and societies is the very reason that individuals need to be able to speak back to culture and reshape it over time. Moreover, the cultural sphere is where individuals find *meaning* in their lives. Culture is a sphere in which individuals share with and seek to understand others. Culture is a sphere that individuals often do not want to leave, or step outside, because culture gives their lives value.[55] At the same time, cultural mores can limit individual freedom, especially when individuals are without sufficient rights to joke about, critique, transgress, and rewrite culture. Culture is both a source of shared meaning and a set of tools for change.[56] Play in culture must include the right to challenge existing culture using the signifiers of that culture itself. The focus on cultural embeddedness does not mean that individualism is lost to the requirements of the community. Rather the idea is that the individual must be understood in context; the individual cannot be stripped from her situation, which is constitutive. At the same time, the individual must have the ability and right to go beyond the limits of her culture and seek to transform it.

Cohen's book is also marked by a concern for particularity that is characteristic of STS. Path dependence, ethnography, and time and place—rather than an abstract search for immanent and universal truths—are all watchwords of STS. Cohen's call of attention to the situated and embedded self in networks of technologies and ideologies requires greater attention to the actual, not theoretical, conditions of creation. Cohen calls for "good story-telling" about how actors create within particular networks.[57]

All of this gives some elaboration to the theoretical insights of cultural theory, especially the theories of STS on which Cohen relies so heavily. But what of the implications of this theory for law, current legal conflicts, and

---

53. *Id.*

54. *Id.*

55. *See generally* Madhavi Sunder, *Cultural Dissent*, 54 STAN. L. REV. 495 (2001) (critiquing expressive association law for forcing members to choose either their culture or their freedom).

56. *Id.* at 498.

57. COHEN, *supra* note 1, at 268. For examples of ethnographies of scientific innovation in STS, see THE SCIENCE STUDIES READER (Mario Biagioli ed., 1999) and ANDREW HARGADON, HOW BREAKTHROUGHS HAPPEN: THE SURPRISING TRUTH ABOUT HOW COMPANIES INNOVATE (2003).

real people in their everyday lives? Because Cohen focuses here on elaborating a theoretical account, this book does not seek to apply it in any detail to current controversies. In contrast to the law and economics model of copyright law, which would justify limitations on author's rights only where there is market failure, a cultural approach would limit rights where they may unduly affect self-actualization. "Autonomy is exercised, and self-determination pursued, by working through culture," Cohen writes.[58] "Laws granting rights in artistic and intellectual expression should be designed with that process in mind."[59] Some of Cohen's concrete suggestions in this regard include advocacy for a "personal use" right that is context sensitive and the reservation of a broad range of remix rights to users.[60]

B. Culture and Capabilities

Culture is a key component of not only individual self-actualization, but human development generally. Cohen, like a growing handful of intellectual property theorists in recent years,[61] turns to the work of Martha Nussbaum and Amartya Sen to flesh out these connections.

> The "capabilities approach" to development pioneered by Amartya Sen and Martha Nussbaum offers a critique of the utilitarian account of development as measured by GDP or technological advancement alone. Sen's vision of "development as freedom" is pluralist, measuring development by assessing an individual's ability to exercise many freedoms, including market-oriented freedom. As Nussbaum has further articulated, central human freedoms range from the fulfillment of basic needs, such as the right to life and health, to more expansive freedoms of movement, creative work, and participation in social, economic, and cultural institutions.[62]

Adopting the capabilities approach (first put forward by an economist, no less!) reaffirms the continuing centrality of economic analysis. At the

---

58. COHEN, *supra* note 1, at 104.

59. *Id.*

60. *Id.* at 246–47.

61. *See supra* note 4.

62. Sunder, *IP*^{3}, *supra* note 4, at 313–14 (footnotes omitted); *see also* NUSSBAUM, WOMEN AND HUMAN DEVELOPMENT, *supra* note 18, at 78–80; *id.* at 5 (defining capability as "what people are actually able to do and to be" in a given society); AMARTYA SEN, DEVELOPMENT AS FREEDOM 3 (1999) [hereinafter SEN, DEVELOPMENT] ("Focusing on human freedoms contrasts with narrower views of development, such as identifying development with the growth of gross national product, or with the rise in personal incomes, or with industrialization, or with technological advance, or with social modernization."); *id.* ("Development can be seen, it is argued here, as a process of expanding the real freedoms that people enjoy."); AMARTYA SEN, INEQUALITY REEXAMINED 37 (1992) (emphasizing "the gap between *resources that help* us to achieve freedom and the extent of *freedom itself*"); Amartya Sen, Equality of What?, The Tanner Lecture on Human Values (May 22, 1979), *available at* http://www.uv.es/~mperezs/intpoleco/Lecturcomp/Distribucion%20Crecimiento/Sen%20Equaliy%20of%20what.pdf (defining "basic capabilities" as "a person being able to do certain basic things").

same time, "the impact of economic growth on human capabilities can be extremely variable, depending on the nature of that growth (for example, how equitable and employment-intensive it is, and whether the economic gains from growth are used to address the deprivations of the most needy.)"[63] Jean Drèze and Sen stress, for example, that "'uncaging' the tiger" of economic development includes the "removal of barriers to using markets," but requires us "to go *well beyond* liberalization."[64] The practical usability of market opportunities, they note, depends on "basic capabilities—including those associated particularly with literacy and education (and also those connected with basic health, social security, gender equality, land rights, local democracy)."[65]

Moving beyond law and economics shifts not only our descriptive landscape, but also the end posts. The normative vision underlying the standard law and economics approach largely embraces wealth as the ultimate value, for practical purposes, if not theoretically elegant ones. While Kaplow and Shavell recognize "the defects in the conceptual and normative foundations of wealth maximization," they believe that "analysis based on wealth maximization" may yet prove "analytically useful."[66] They offer the same defense with respect to "efficiency."[67]

The human capabilities approach on which Cohen bases her work has a different goal in mind. Martha Nussbaum, one of the principal architects of this approach, reminds us of its origins, when philosophers and developmental economists stopped to "[s]uppose for a moment that [they] were interested not in economic or political theory but just in people."[68] As the late Mahbub ul Haq, the mastermind behind the U.N. Human Development Reports, explained in the first such report in 1990: "People are the real wealth of a nation. The basic objective of development is to create an enabling environment for people to enjoy long, healthy and creative lives."[69]

A copyright law grounded in the capabilities approach, in contrast to traditional law and economics analysis, ought to focus then on more than the creation of more goods. We need to measure law's success by its ability to better the lives of real people. In short, a cultural turn in intellectual property provides new answers to the fundamental question: What is intellectual property for?

---

63. JEAN DRÈZE & AMARTYA SEN, INDIA: DEVELOPMENT AND PARTICIPATION 37 (2002).

64. *Id.* at 308.

65. *Id.*

66. Louis Kaplow & Steven Shavell, *Fairness Versus Welfare*, 114 HARV. L. REV. 961, 997 (2001).

67. *Id.*

68. NUSSBAUM, *supra* note 14, at 3–4.

69. UNITED NATIONS DEVELOPMENT PROGRAMME, HUMAN DEVELOPMENT REPORT 1990, at 9 (1990).

Seeking to set out a broad theoretical account, Cohen does not herself seek to offer examples of how her theory affects real people in real situations. This makes it harder to figure out how the book's arguments might work out in practical form. In particular, it is important to understand the need for particularity in the context of an intellectual property law that has become globalized. Because of TRIPS, American intellectual property scholars must increasingly consider contexts outside the United States.[70] We must thus address contemporary issues in international copyright, from access to copyrighted materials for the disabled and the poor, to how to enhance the ability of peoples around the world to create their own knowledge of the world.

Which brings us back to Vasanti. We suggest that a cultural approach to copyright premised on the capabilities approach needs to attend to how copyright law can expand her capabilities. Vasanti is illiterate (or at least was so at the time of Nussbaum's writing), so she will perhaps be keen on educational texts that might help her learn to read. What is Vasanti's ability to access educational works? What about Vasanti's access to popular works that comprise a common, cross-cultural lexicon like J.K. Rowling's *Harry Potter* series? Market theorists are content with whatever culture the market produces, paying no attention to who produces it, who can access it, or for whom it is written. The new cultural theorists begin with a deep engagement with culture and build the theory from there. Culture gives us a common vocabulary, a shared set of experiences on which to build. If Vasanti is unable to access works that the whole world knows[71] she may be excluded from cross-cultural discourse.[72]

Does someone like Vasanti find time for play—described by Nussbaum as one of the ten basic capabilities?[73] Perhaps Vasanti enjoys Bollywood films. Some of the works that Vasanti might learn from or enjoy may be priced out of reach or unavailable in her vernacular language, Gujarati. If she enjoys big-budget film productions, Vasanti might want a copyright law that enables producers to invest capital into a film and earn a reasonable rate of return. She might also want to enjoy rights to critique the work. Vasanti must be engaged in creating her world. She might well want the ability to speak back, through cultural works themselves—perhaps to criticize many Bollywood films for their disproportionate attention to the lives of the very

70. The World Trade Organization's Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) imposes minimum standards of intellectual property protection and enforcement on member nations. WORLD TRADE ORG., UNDERSTANDING THE WTO 39 (5th ed. 2011).

71. SUNDER, FROM GOODS TO A GOOD LIFE, *supra* note 4, at 94.

72. *See* NUSSBAUM, *supra* note 14, at 7 (describing how the lack of access to historical, economic, political, and literary works can cut an individual off from a full understanding of a culture).

73. *Id.* at 34.

rich or their depiction of women like her. As Foucault famously noted, culture can be disciplinary, seeking to confine freedom and identities.[74] So too is culture a tool for reform; cultural revolution results from people seeking to transgress their cultural boundaries.

Vasanti is part of a global copyright order where any creative work automatically receives copyright protections across most of the world. Perhaps there should be market segmentation, allowing for cheaper or even free works to be made available to Vasanti—perhaps through labeling—and more expensive versions to wealthier individuals.

Cohen is concerned with flexibilities in new technologies that allow her to manipulate cultural works—are these technologies available to Vasanti? Vasanti's access to capability-enhancing tools (from libraries to the Internet) in turn affects her capacity to create and contribute to our global cultural heritage. Her potential to be a creator of intellectual works, themselves protected by fair intellectual property laws, might even offer economic value for her and her family.

C. *Objections*

There are two principal objections to the idea of expanding our methodological inquiry beyond economics and our normative vision beyond incentivizing creativity. The worry about methodological pluralism is that it complicates the analysis too much to be useful. A similar worry attends normative pluralism—but with an additional concern that to entertain values other than efficiency is to authorize the dramatic expansion of intellectual property rights. We consider these concerns here.

Too Complex. The elegant simplicity of the syllogism that more property rights yields more creativity,[75] however, leads to substantial error. Cohen worries about "legal scholars' reluctance to engage culture in its own right, without the filters supplied by simplistic economic models or by more complex models derived from the life sciences."[76] Recognizing the myriad of variables involved in the creative process will expand our intellectual creativity policy making beyond simply the copyright term and scope to involve issues such as the creative environment, the availability of existing cultural work for commentary and manipulation, and the freedom of

74. *See generally* MICHEL FOUCAULT, DISCIPLINE & PUNISH: THE BIRTH OF THE PRISON (trans. Alan Sheridan, Vintage Books, 2d ed. 1995) (1975) (constructing a theory of society that links the creation of the modern penal system to the rise of enlightenment thinking and illustrating that the existence of society is inherently disciplinary).

75. We recognize that many scholars have pointed out the economic value of the public domain, so economic analysis does not necessarily lead to maximalist copyright. Yet, important strains of the economics approach still exhort more protection. More fundamentally, even recognizing the public domain's economic value is not sufficient, as we argue. Both copyrighted works and the public domain have non-economic value.

76. COHEN, *supra* note 1, at 24.

expression. As Cohen points out, we create within physical environments that interact with our intellectual energies in often unexpected ways.

It is indeed daunting to consider the bewildering complexity and irrationality of culture, history, and psychology. Yet, it only makes sense to economize on theory if the results will not be too far off, if the omitted details are largely unimportant. Paul Krugman describes this as "mistaking beauty for truth."[77] As we have written elsewhere, "[I]f our move [to add additional values to intellectual property decision making] adds complexity, it is just the complexity necessary to get things right. *Narrowing the calculus to ease the calculation will likely lead to the wrong answer.*"[78] As John Law writes, "If this is an awful mess . . . then would something less messy make a mess of describing it?"[79]

Too Much Intellectual Property. Many liberal theorists of intellectual property rights worry that to entertain values other than efficiency is to authorize the dramatic expansion of intellectual property rights. We believe that the additional normative concerns provide resources to limit that expansion. The economic rationale counsels nearly boundless expansion as long as it can be justified by some (even implausible) claim that more property rights induce more creativity. A pluralist account of intellectual property might counsel restraint in expanding rights. Consider a real world instance of this: in England, an appeals court "relied on human rights law to establish a compulsory license allowing a paper to publish a memo of a secret meeting with Prime Minister Tony Blair, despite claims that it would infringe copyright."[80]

Conclusion

How should we think about the domain of human life subject to copyright? Should we focus exclusively on information and other transaction costs, free riding, optimum terms, and remuneration? Or should we include concerns such as inspiration, desire, emotion, predicament, necessity, anger, joy, hunger, ennui, anomie, network, bodies, children, death, play, and love?

Cohen's book marks a major effort to expand the vocabulary and concepts of intellectual property. We celebrate this effort. Scholars should seek to make intellectual property law more human.

---

77. Paul Krugman, *How Did Economists Get It So Wrong?*, N.Y. TIMES, Sept. 2, 2009, http://www.nytimes.com/2009/09/06/magazine/06Economic-t.html.

78. Anupam Chander & Madhavi Sunder, *Is Nozick Kicking Rawls's Ass? Intellectual Property and Social Justice*, 40 U.C. DAVIS L. REV. 563, 577 (2007).

79. JOHN LAW, AFTER METHOD: MESS IN SOCIAL SCIENCE RESEARCH 1 (2004).

80. Chander & Sunder, *supra* note 78, at 578.

Purposive Hopes for Better IP

CREATION WITHOUT RESTRAINT: PROMOTING LIBERTY AND RIVALRY IN INNOVATION. By Christina Bohannan & Herbert Hovenkamp. New York, New York: Oxford University Press, 2012. 440 pages. $45.00.

Reviewed by John M. Golden*

In *Creation Without Restraint*,[1] Christina Bohannan and Herbert Hovenkamp add to the growing body of books on what academics have commonly come to call a "crisis"—namely, the current state of United States' intellectual property (IP) laws and their interaction with policies of promoting both innovation and free-market competition. Bohannan and Hovenkamp embrace the terminology of "crisis,"[2] but more fundamentally focus on a series of specific problems with how our modern patent, copyright, and antitrust laws operate and how their operation might be improved. Although they might not have hit on a cure-all, their diagnoses and proposed cocktail of reforms are well worth considering regardless of whether one accepts the crisis terminology. At least Justice Stephen Breyer of the United States Supreme Court seems to agree. Through multiple citations of *Creation Without Restraint* in an opinion for the Court he authored in *Mayo Collaborative Services v. Prometheus Laboratories, Inc.*,[3] Bohannan and Hovenkamp's book has already made a notable appearance in public debate.[4]

The breadth of Bohannan and Hovenkamp's project impresses. They do not openly confine themselves to any particular type of innovation and define the term "innovation" broadly to encompass "any human idea that adds something important to what we already have."[5] Further, the authors do not confine themselves to analyzing how one particular form of government

* Professor in Law, The University of Texas at Austin. I thank Oren Bracha for comments on a draft version of this Review, and I thank the editors of the *Texas Law Review* for their help in bringing this Review to its final published form.

1. CHRISTINA BOHANNAN & HERBERT HOVENKAMP, CREATION WITHOUT RESTRAINT: PROMOTING LIBERTY AND RIVALRY IN INNOVATION (2012).

2. *Id.* at xiv ("The patent system is in a crisis of overissuance, overprotection, and excessive litigation. . . . The future is bleaker for copyright law."); *id.* at 60 ("Today the U.S. patent system is in crisis."); *id.* at 133 ("The crisis in copyright law today is just as serious as the one in patent law. . . .").

3. 132 S. Ct. 1289 (2012).

4. *See id.* at 1302 (citing Bohannan and Hovenkamp's book in support of the notion that overly broad patent rights can slow innovation); *id.* at 1305 (citing Bohannan and Hovenkamp's book in support of the proposition that "the practical effects of [patent law's] rules . . . may differ from one field to another").

5. BOHANNAN & HOVENKAMP, *supra* note 1, at ix.

action helps foster or impede innovation. Instead, the authors substantially take on at least three such regimes: modern patent, copyright, and antitrust laws. In some ways, they might be criticized for taking on too much at too fine a level of granularity: as Robert Merges has recently observed, present-day IP law by itself "is like one of those sprawling, chaotic megacities of the developing world"[6]—diverse, protean, and resistant of uniformly firm handles. Fortunately, the authors can draw on a depth of experience and knowledge in analyzing questions relating to IP and antitrust topics.[7] The result is a combination of information and thought that should enrich the understanding of any reader.

Needless to say, a brief review cannot hope to do justice to such a book's contents.[8] I start with a quick overview followed by a sampling of some of the book's more detailed contents.

The book has thirteen numbered chapters plus an introduction and epilogue. To my eye these fifteen subdivisions coalesce into essentially five parts. The first runs from the introduction through Chapter 3. This part presents relatively general thoughts about intellectual property, antitrust, and the workings of markets[9] and concludes by arguing generally for a requirement of cognizable "IP injury" before the awarding of remedies for IP infringement.[10] The second part, Chapters 4 and 5, focuses on the patent system, denies that antitrust "offer[s] a global fix" to its problems,[11] and suggests a variety of reforms. The third part, Chapters 6 through 8, turns to

---

6. ROBERT P. MERGES, JUSTIFYING INTELLECTUAL PROPERTY 1 (2011).

7. *See, e.g.*, Christina Bohannan, *Copyright Harm, Foreseeability, and Fair Use*, 85 WASH. U. L. REV. 969, 969 (2007) (contending that "[c]opyright law needs a theory of harm that can give effect to its constitutional purpose"); Christina Bohannan, *Copyright Infringement and Harmless Speech*, 61 HASTINGS L.J. 1083, 1087 (2010) (arguing that, like other forms of speech regulation, copyright restrictions on speech must be "necessary to prevent or remedy harm to a sufficient government interest"); Christina Bohannan & Herbert Hovenkamp, *IP and Antitrust: Reformation and Harm*, 51 B.C. L. REV. 905, 907–08 (2010) ("[U]rg[ing] courts to develop the concept of 'IP injury,' similar to the concept of 'antitrust injury' in the antitrust laws, which links the type of harm that a plaintiff must show to the underlying purpose of those laws"); Christina Bohannan, *IP Misuse as Foreclosure*, 96 IOWA L. REV. 475, 478 (2011) (seeking to redirect doctrines of IP misuse to the question of "whether an alleged act of misuse violates IP policies of encouraging innovation, promoting competition, or encouraging access to the public domain"); Herbert Hovenkamp, *Antitrust and Innovation: Where We Are and Where We Should Be Going*, 77 ANTITRUST L.J. 749, 750 (2011) ("offer[ing] a few principles for antitrust analysis in innovation-intensive markets"); Herbert Hovenkamp, *Patents, Property, and Competition Policy*, 34 J. CORP. L. 1243, 1243 (2009) (discussing whether "competition policy [should] have a more prominent role than it currently has in helping the patent system promote innovation"); Herbert Hovenkamp, *Restraints on Innovation*, 29 CARDOZO L. REV. 247, 260 (2007) (contending that "[r]estraints on innovation deserve the special attention of the government agencies charged with enforcing the antitrust laws").

8. For another effort, see Paul R. Gugliuzza, *IP Injury and the Institutions of Patent Law*, 98 IOWA L. REV. 747 (2013) (reviewing BOHANNAN & HOVENKAMP, *supra* note 1).

9. *See, e.g.*, BOHANNAN & HOVENKAMP, *supra* note 1, at 17 ("Many competition disputes in technology-rich markets involve claims about interconnection, compatibility, or interoperability.").

10. *Id.* at 51 ("IP law should recognize harm only for uses that are likely to interfere with IP holders' decisions to create or distribute their works . . . .").

11. *Id.* at 97.

copyright. This part describes how, to the apparent detriment of social welfare and free speech interests, the Copyright Act has come to "favor rent-seeking special-interest groups rather than the general public,"[12] and then outlines how courts might interpret the Constitution and Copyright Act "to reclaim copyright law for the public interest."[13] The fourth part, Chapters 9 through 13, returns to more general concerns with competition, innovation-supporting policy, and various forms of "commons" or "semicommons" for the sharing of information resources.[14] Finally, the book's fifth and shortest part—the "Epilogue"—provides a summary list of eleven reform proposals that draw on prior discussions.

What are some of the book's more specific contents? I think four contentions from the book's first part are worth particular emphasis:

1. much of the solution to modern problems with IP must come through reform of IP laws or their understanding, rather than through more vigorous or creative antitrust enforcement;[15]
2. would-be IP reformers can learn much from antitrust's evolution toward an economics-oriented regime with a requirement of "antitrust injury" for private suits,[16] namely, a requirement that a private plaintiff bringing an antitrust suit allege "not just any injury, but *antitrust* injury—that is, injury that results from decreased competition";[17]
3. more specifically, the IP laws should possess a requirement of "IP injury" for private-enforcement suits;[18] and
4. "courts are more likely than Congress to be the engines of significant reform" in the operation of IP law.[19]

A starting point for the authors' contentions is their view that IP laws, like antitrust laws designed "to promote competition,"[20] are "regulatory regimes" with a purpose.[21] Pointing to the U.S. Constitution's "IP clause," which demands "that patent and copyright law 'promote the progress' of their fields by creating rights for 'limited times,'" the authors work on the basis of an assumption that the "rationale for IP" comes from its capacity to

12. *Id.* at 136.

13. *Id.* at 200.

14. *Id.* at 325–26.

15. *See id.* at 12 ("Many competition issues can be addressed more effectively through the IP statutes themselves, either alone or in addition to antitrust law."); *id.* at 13 ("[I]t is not antitrust's purpose to fix political defects or to cure shortcomings in other regulatory regimes.").

16. *Id.* at 15 ("IP law can take some important lessons from the road that antitrust law has taken toward reform.").

17. *Id.* at 49.

18. *Id.* at 51.

19. *Id.* at 34.

20. *Id.* at xi.

21. *Id.* at 13 (characterizing IP among a number of "regulatory regimes" with which antitrust law interacts); *see also id.* at 45 ("[T]he IP laws are affirmative regulatory provisions . . . .").

advance social-welfare-improving innovation by providing "economic incentives rather than some alternative theory such as natural rights."[22] In light of this understanding, the authors assert that the fundamental problems with current IP law are much like prior problems in antitrust law—a disconnect between law and law's "articulated goals" and a failure to appreciate that markets "actually work[] much better on their own," without "government intervention," than IP or antitrust advocates have commonly suggested.[23]

At least if one puts aside concerns of administrability, the sort of "IP injury" requirement that Bohannan and Hovenkamp propose seems a reasonable response to a desire to bring IP law more into conformity with its goals. Bohannan and Hovenkamp's basic formulation of the IP injury requirement is that it would demand a showing of harm that is "likely to interfere with IP holders' decisions to create or distribute their works"—i.e., "demonstrable injury that is tied to the purpose for which the IP laws were passed in the first place."[24] In at least partial answer to those who might worry that this basic formulation is too uncertain in scope,[25] Bohannan and Hovenkamp provide helpful albeit not entirely definitive elaborations on what they conceive to be IP injury toward the end of the book's first part and as part of their efforts to delineate "copyright harm" in the book's third part.[26] For example, they explain that IP injury would occur in situations where infringement "clearly deprives the rights holder of sales"[27] or where "it is clear that an innovator would rely on [sought-after] royalties in deciding whether to create the work."[28] But a demand for IP injury "would suggest little or no protection for situations where any harm caused by the alleged infringement is merely speculative"—for example, "where the defendant produces a complementary work that increases sales of the [less ambiguously] protected work, or in which the defendant uses the work for

22. *Id.* at x; *see also id.* at 46 ("About the best we can say is that the primary goal of IP policy should be to maximize net gains from innovation after all transaction costs have been paid."). *But see* MERGES, *supra* note 6, at 15 ("The basic foundations of IP law are individual autonomy and freedom.").

23. BOHANNAN & HOVENKAMP, *supra* note 1, at 35.

24. *Id.* at 51.

25. *See, e.g.*, Shyamkrishna Balganesh, *Foreseeability and Copyright Incentives*, 122 HARV. L. REV. 1569, 1606 (2009) (preferring a foreseeability requirement for copyright-infringement liability that "focus[es] on the defendant's actions (that is, the copying), rather than function[ing] as an open-ended device that courts might then connect to the notions of 'harm' or 'market,'" concepts that could introduce "[q]uestions of appropriate baselines, market substitutability, remoteness, and the like").

26. BOHANNAN & HOVENKAMP, *supra* note 1, at 176–99 (describing situations in which harm might "be presumed" and various factors that courts should consider in determining whether there has been "copyright harm").

27. *Id.* at 57.

28. *Id.* at 58.

personal, noncommercial purposes for which people would ordinarily be unwilling to pay."[29]

As a potential example of a noninjurious complementary work, they cite the fictional book *The Da Vinci Code*, "a religious historical thriller involving a romantic relationship that allegedly existed between Jesus Christ and Mary Magdelene."[30] Publication of *The Da Vinci Code* apparently helped generate greatly increased sales of a nonfiction book *Holy Blood, Holy Grail*, "which explored the Jesus and Mary Magdalene story."[31] Nonetheless, the publisher of the latter book was less than grateful, and an infringement suit against the publisher of *The Da Vinci Code* followed.[32] Bohannan and Hovenkamp suggest that, under an IP injury doctrine, such a suit might be quickly and efficiently dismissed.[33]

At times, Bohannan and Hovenkamp seem to waver on whether Congress should be expected to be involved in implementing a requirement of IP injury.[34] Perhaps this is in part because, from their perspective, the current situation in copyright law is so grim that reform might not be reasonably expected even from the courts.[35] Generally speaking, however, Bohannan and Hovenkamp suggest that the courts are the most likely and reliable agents of reform.[36] In their view, lawmakers are too liable to special-interest capture.[37] Courts are better insulated from special-interest pressures and thus more capable of bringing IP laws back into accord with their

29. *Id.*; *see also id.* at 183 ("Proof of harm must also take into account the positive effects of unauthorized copying.").

30. *Id.* at 53.

31. *Id.*

32. *Id.*; *see also id.* at 183 (describing a report of increased sales of *Holy Blood, Holy Grail* after *The Da Vinci Code*'s release).

33. *Id.* at 55.

34. *Compare id.* at 59 ("In developing an IP injury or harm requirement, courts should keep in mind that the patent and copyright laws have explicit authorization in the Constitution . . . . Thus, IP has a powerful guiding principle. It need only be used."), *and id.* at 199 ("Both the IP Clause and the First Amendment require proof of harm for copyright infringement liability."), *with id.* at 51 ("As a first step in their own reform journeys, drafters of the IP laws need to develop a more disciplined conception of 'IP injury.'"), *and id.* at 181 ("Both Congress and the courts should recognize that the touchstone of infringement is harm to the copyright holder's incentives.").

35. *See id.* at 47 ("[S]ome recent judicial decisions and pending patent reform legislation show that patent law has begun its own reform journey . . . . The outlook for copyright law is bleaker."); *id.* at 237 ("The reality is . . . that the Supreme Court has granted copyright law a measure of constitutional 'exceptionalism' that severely limits the range of permissible constitutional attacks on copyright overreaching.").

36. *See, e.g.*, *id.* at 15 (noting that development of an "'antitrust injury' doctrine" "was accomplished entirely by federal judges, largely in the face of congressional indifference and in apparent conflict with a private injury statute that guarantees liberal recovery for every kind of injury"); *id.* at 34 ("[C]ourts are more likely than Congress to be the engines of significant reform.").

37. *See id.* at 47 ("The classic public choice paradigm clearly favors IP rights holders . . . .").

constitutionally mandated purpose[38] "[t]o promote the Progress of Science and useful Arts."[39]

What else do Bohannan and Hovenkamp tell us? There is a lot else—too much to discuss in real detail here. So I will offer only a few additional tidbits from the book's second through fourth parts, which precede the fifth part's summary of proposals.

The book's second and third parts offer a host of specific suggestions for reforming patent and copyright, respectively. For example, Bohannan and Hovenkamp contend that patent claims "not included in the original application" should be either entirely unenforceable or at least unenforceable "retroactively against those who made a technological choice before the claim was on record."[40] Under certain circumstances, a so-called reverse payment settlement, in which a patentee pays an accused infringer to settle an infringement suit, "might remove the presumption of patent validity" or automatically "trigger reexamination" of the patent by the United States Patent and Trademark Office.[41] Courts should enable more frequent and easier rejection of patent claims on grounds that the claimed invention is obvious relative to "common knowledge about how the world works."[42] Congress should enact a general independent-invention defense against patent infringement.[43] Just as continued patent protection requires periodic payment of maintenance fees, copyright protection should require periodic renewal.[44] Courts should more narrowly construe the scope of derivative works covered by copyright.[45] I could continue with such examples or their elaboration, but space restrictions do not allow it.

Instead, I'll proceed to give a taste of the book's fourth part, which has a more competition-oriented focus. Echoing the book's title, Bohannan and Hovenkamp forcefully argue in Chapter 9 that antitrust should show more active concern about "[r]estraints on innovation."[46] Bohannan and Hovenkamp point out that the social stakes appear high because "[t]oday

38. *See id.* at 395 ("[C]ourts are freer from interest-group pressures and thus in a better position to make wise decisions.").

39. U.S. CONST. art. I, § 8, cl. 8.

40. BOHANNAN & HOVENKAMP, *supra* note 1, at 75.

41. *Id.* at 95 (describing potential results of "a high exit payment" by the patentee).

42. *Id.* at 110.

43. *Id.* at 128 ("[T]he case for a suitably constrained right of independent invention . . . is so strong that it merits discussion.").

44. *Id.* at 202 (stating that "the principle [of requiring periodic renewal of copyright] is a good one").

45. *Id.* at 223–24 ("Copyright holders should be entitled to control the markets for the forms of their works listed in the statutory definition of derivative works, but not the markets for works with substantially new content or purpose.").

46. *Id.* at 238; *see also id.* at 245 ("One place the antitrust laws could be more aggressive then than they are today is . . . . in policing practices that restrain the innovations of others without a serious and provable efficiency-related explanation."); *id.* at 251 (stating that antitrust "enforcement [against innovation restraints] has never been as strong as it should be").

there is little doubt that innovation contributes far more to economic growth than does the movement of markets from less to greater amounts of price competition"[47]—the latter being a more traditional obsession of antitrust. Chapter 11 outlines some specific ways in which antitrust enforcement can work to facilitate innovation. For example, antitrust law could episodically "impose sharing obligations in dominated networks"[48] or condemn extreme cases of "'predatory' product innovation"[49] such as that seeking to exclude rivals through a "technological ti[e]" like "Microsoft's 'commingling' of Windows and [Internet Explorer] code."[50]

Although Bohannan and Hovenkamp propose such reforms, they also emphasize that antitrust is no panacea. In their view, administrability concerns place fundamental restraints on the capacity of antitrust to "free" innovation: proving a practice to be sufficiently anticompetitive to be forbidden by antitrust is frequently difficult, and calculating damages from such a practice tends to be even harder.[51] As Chapter 12 highlights, different forms of information-based commons or semi-commons can, as in the case of technological standards or patent pools, naturally involve coordination or sharing by competitors. Antitrust can have special difficulty distinguishing between when such activities are socially helpful and when they are socially harmful.[52] Consequently, Bohannan and Hovenkamp look to IP laws themselves for means to address "restraints on innovation," and they find one answer in the form of revitalized doctrines of patent and copyright misuse.[53] On the other hand, Bohannan and Hovenkamp argue in Chapter 13 that the U.S. Supreme Court recently went too far in restricting the effectiveness of post-sale restraints on use of a patented invention.[54] Bohannan and Hovenkamp contend that, rather than a per se rule limiting the effectiveness of such restraints, courts should use "[r]ule-of-reason analysis under the antitrust laws or perhaps patent misuse doctrine" to distinguish "between the harmful and the harmless."[55]

---

47. *Id.* at 239; *cf.* ROBERT D. COOTER & HANS-BERD SCHÄFER, SOLOMON'S KNOT: HOW LAW CAN END THE POVERTY OF NATIONS ix (2012) ("The central claim of this book is that sustained growth in developing countries occurs through innovations in markets and organizations by entrepreneurs . . . .").

48. BOHANNAN & HOVENKAMP, *supra* note 1, at 318.

49. *Id.* at 322.

50. *Id.* at 320–21; *see also id.* at 323 ("[C]laims of anticompetitive product innovation should be limited to the very small number of situations where it is clear from the outset that the dominant firm was not attempting to improve its own product, or at least not more than trivially, but only to injure the market of a rival.").

51. *Id.* at 254 (discussing administrability concerns).

52. *Id.* at 363 (noting difficulties in "evaluat[ing] practices that involve competitor sharing, such as pooling or standard setting").

53. *Id.* at 256.

54. *Id.* at 389 ("[T]he first-sale rule, which operates as a per se restraint, seems excessive . . . .").

55. *Id.*

This discussion of the proper scope of patent law's first-sale doctrine underlines the fundamental pragmatism of Bohannan and Hovenkamp's purposivist approach. They are not dogmatic in their search for better ways to achieve stated social goals. Sometimes, as with the first-sale doctrine, they argue that their approach calls for loosening limitations on IP rights. Many times, unsurprisingly in light of their general perception of IP overreach,[56] they argue in the opposite direction.

In a book that provides so much, where would I have liked to have seen more? Perhaps my greatest want would be for a more extended and forceful argument for the purposive approach to statutory and constitutional interpretation that Bohannan and Hovenkamp appear almost to take for granted. As inspiration for many of their proposals, Bohannan and Hovenkamp point to a revolution in antitrust law that gathered legal force in the 1970s.[57] Within this antitrust revolution, they particularly highlight the U.S. Supreme Court's 1977 decision in *Brunswick Corp. v. Pueblo Bowl-O-Mat, Inc.*[58] Bohannan and Hovenkamp characterize *Brunswick*'s embrace of an antitrust injury requirement for private suits as "defy[ing] the clear language" of the antitrust statute[59] by "impos[ing] a limiting interpretation on a private enforcement provision that seems clear and expansive on its face."[60]

Bohannan and Hovenkamp suggest that the purposive moves that they advocate today, including judicial embrace of an "IP injury" requirement, fall far short of demanding the boldness of *Brunswick*. In their words, "In patent and copyright law, . . . the courts need not defy the clear language of the statute as the Supreme Court did in the case of antitrust."[61]

I am not sure that all reasonable minds will agree that an "IP injury" requirement does not run afoul of plain statutory language. For example, the U.S. Patent Act contains statutory language providing, "A patentee shall have remedy by civil action for infringement of his patent."[62] I can envision a textualist argument that this language uses an unequivocal "shall" to give patentees an entitlement to a civil remedy that is not conditioned on any "IP injury" other than "infringement," a concept that § 271 of the Patent Act, entitled "Infringement of Patents," separately defines.[63]

---

56. *Id.* at 404 ("In the great IP battle of appropriation versus access, today appropriation has the upper hand.").

57. *See id.* at 34–39 (discussing "[t]he story of antitrust reform" as an instructive example for potential IP reform).

58. 429 U.S. 477 (1977).

59. BOHANNAN & HOVENKAMP, *supra* note 1, at 34.

60. *Id.* at 48; *see also id.* at 50 ("*Brunswick*'s most notable feature is its virtual disregard of the language of antitrust's statutory private action provision . . . .").

61. *Id.* at 34.

62. 35 U.S.C. § 281 (2006).

63. *Id.* § 271; *see also, e.g.*, *id.* § 271(a) ("Except as otherwise provided in this title, whoever without authority makes, uses, offers to sell, or sells any patented invention, within the United States or imports into the United States any patented invention during the term of the patent therefor, infringes the patent.").

In any event, even if Bohannan and Hovenkamp are right that textual barriers to their proposed reforms are inherently weaker than those overcome by the Supreme Court in *Brunswick*, the relevant legal environment has changed dramatically since 1977. The result might be that *inherently weaker* textual barriers are *circumstantially stronger*.

It is not news that a textualist revolution has swept across the U.S. legal world since the mid-1980s.[64] Although advocates of purposivist interpretation such as Justice Stephen Breyer continue to argue the superiority of their approach,[65] even they give evidence of a felt need to fit their interpretations "within the semantic boundaries of the text."[66] Hence, although discovering an IP injury requirement in patent and copyright law might be a theoretically easier task than discovering an antitrust injury requirement in federal antitrust laws, the former discovery might in fact be much less likely in the present "new textualist" world.[67]

IP law has been far from immune from the new textualism. Indeed, recent decisions by the U.S. Supreme Court tend to suggest that, for the foreseeable future, textualism will significantly constrain the likelihood of judicially based reform along the lines that Bohannan and Hovenkamp advocate. In the Supreme Court's 2010 opinion in *Bilski v. Kappos*,[68] Justice Kennedy, writing for a majority of five justices, stressed fidelity to the "ordinary meaning" of statutory text as a basis for rejecting a requirement that a patentable process involve "a particular machine or apparatus" or "transfor[m] a particular article into a different state or thing."[69] In so doing, the Court seemed to draw a line in the sand with respect to nontextualist glosses on § 101 of the U.S. Patent Act.[70] This statutory provision describes the types of things—a "process, machine, manufacture, or composition of matter"—that a utility patent may cover.[71] The Court acknowledged that its "precedents provide three specific exceptions to § 101's broad patent-

64. JOHN F. MANNING & MATTHEW C. STEPHENSON, LEGISLATION AND REGULATION 67 (2010) ("Over the last quarter-century, textualism has had an extraordinary influence on how federal courts approach questions of statutory interpretation."); *see also* William N. Eskridge, Jr., *The New Textualism*, 37 UCLA L. REV. 621, 624 (1990) ("The new textualism is the most interesting development in the Court's legisprudence (the jurisprudence of legislation) in the 1980s . . . ."); Victoria Nourse, *Misunderstanding Congress: Statutory Interpretation, the Supermajoritarian Difficulty, and the Separation of Powers*, 99 GEO. L.J. 1119, 1136 (2011) (observing that in 1987 Justice Scalia provided an account of "the 'new textualism'" that "influenced a generation of legal scholars").

65. *See, e.g.*, STEPHEN BREYER, MAKING OUR DEMOCRACY WORK: A JUDGE'S VIEW 94 (2010) (arguing in favor of "a purpose-oriented approach" to statutory interpretation).

66. MANNING & STEPHENSON, *supra* note 64, at 78 (citing opinions written by Justice Souter and Justice Breyer, respectively, as examples).

67. *See id.* at 49 (describing the rise of "new textualism").

68. 130 S. Ct. 3218 (2010).

69. *Id.* at 3225–26 (internal quotation marks omitted).

70. *See id.* at 3231 ("Today, the Court once again declines to impose limitations on the Patent Act that are inconsistent with the Act's text.").

71. 35 U.S.C. § 101 (2006).

eligibility principles: laws of nature, physical phenomena, and abstract ideas."[72] But the Court cautioned that toleration of these exceptions' continued existence should not invite the imposition of additional restrictions on § 101's explicit scope.[73] Specifically, the Court emphasized that it had "more than once cautioned that courts should not read into the patent laws limitations and conditions which the legislature has not expressed."[74] Unless the Act itself provided contrary definitions, its terms should be "interpreted as taking their ordinary, contemporary, common meaning."[75]

In light of the above principles, the Supreme Court "once again decline[d] to impose limitations on the Patent Act that are inconsistent with the Act's text."[76] More particularly, the Court rejected a restrictive reading of the § 101 term "process" that the U.S. Court of Appeals for the Federal Circuit had adopted—namely, an understanding that a "process" was eligible for patent protection under § 101 "only if: (1) it is tied to a particular machine or apparatus, or (2) it transforms a particular article into a different state or thing."[77] The Supreme Court likewise rejected an argument by justices who concurred in the judgment that "business methods are not patentable."[78] The Court majority observed that "[t]he term 'method,' which is within [the Patent Act's] definition of 'process,' at least as a textual matter . . . , may include at least some methods of doing business."[79] The majority was unmoved by Justice Stevens's contentions for the concurrers that the Court's textualist approach was "deeply flawed" and would generate "absurd results" not only in the patent context at hand[80] but also if applied to interpretation of federal antitrust law.[81]

Bilski is not an isolated example of textualism's influence on IP law. In *Golan v. Holder*,[82] the U.S. Supreme Court rejected an opportunity to embrace a substantially nontextual limitation on U.S. copyright law.[83] The Court also rejected the limiting notion, championed by Bohannan and

72. *Bilski*, 130 S. Ct. at 3225 (internal quotation marks omitted).

73. *Id.* at 3226 ("This Court has not indicated that the existence of these well-established exceptions gives the Judiciary *carte blanche* to impose other limitations that are inconsistent with the text and the statute's purpose and design.").

74. *Id.* (internal quotation marks omitted).

75. *Id.* (internal quotation marks omitted).

76. *Id.* at 3231.

77. *Id.* at 3225 (internal quotation marks omitted).

78. *Id.* at 3232 (Stevens, J., concurring in the judgment).

79. *Id.* at 3228 (opinion for the Court).

80. *Id.* at 3238 & n.5 (Stevens, J., concurring in the judgment).

81. *Id.* at 3238 n.4 ("[I]f this Court were to interpret the Sherman Act according to the Act's plain text, it could prohibit the entire body of private contract." (internal quotation marks omitted)).

82. 132 S. Ct. 873 (2012).

83. *See id.* at 888 (responding to an argument that copyright legislation must stimulate "[t]he creation of at least one new work" by observing that "[n]othing in the text of the [U.S. Constitution's] Copyright Clause confines the 'Progress of Science' exclusively to 'incentives for creation'").

Hovenkamp's book, that "the goal of copyright is to encourage the production of creative works."[84]

In *Golan*, petitioners challenged a statutory provision that purportedly helped bring the United States into compliance with the Berne Convention by extending copyright protection to certain foreign works that had previously been denied copyright protection in the United States.[85] According to an opinion for the Court authored by Justice Ginsburg, "Nothing in the text of the Copyright Clause confines the 'Progress of Science' exclusively to 'incentives for creation,'" and "[e]vidence from the founding . . . suggests that inducing *dissemination*—as opposed to creation—was viewed as an appropriate means to promote science" in accordance with the constitutional charge.[86] The Court also cited precedent to support its lack of belief in the decisiveness of a concern for which Bohannan and Hovenkamp indicate sympathy[87]—namely, that the extension of copyright protection to already existing works, previously in the public domain, does little, if anything, to stimulate the production of new works.[88] Nonetheless, the Court stressed, "Even were [it] writing on a clean slate, petitioners' argument [that copyright legislation must stimulate generation of new work] would be unavailing."[89] As in *Bilski*, the Court rooted in plain language its rejection of a less robust view of IP law: the Court began its rejection of the petitioners' constitutional challenge by stating, "The text of the Copyright Clause does not exclude application of copyright protection to works in the public domain."[90]

The opinions in *Bilski* and *Golan*, penned by Justices who are far from the strongest champions of textualism, leave me with little proximate hope that either the Supreme Court or lower courts will embrace something like Bohannan and Hovenkamp's IP injury requirement in any truly robust form. Indeed, the opinions suggest to me that Bohannan and Hovenkamp's general enterprise, calling for more sensibly purposivist interpretation of the IP laws, faces a more uphill battle than they appear openly to acknowledge. Because of my perception of the array of opposing forces, I would have liked to see them make more forceful arguments for the purposivist approach to interpretation for which they sometimes seem almost to presume acceptance.

---

84. BOHANNAN & HOVENKAMP, *supra* note 1, at 204; *see also id.* at 59 (observing that the constitutional "authorization [for patent and copyright] expressly ties the IP rights created to the incentive to create").

85. 132 S. Ct. at 877–78 (describing the contested provision of the 1994 Uruguay Round Agreements Act).

86. *Id.* at 888.

87. *See* BOHANNAN & HOVENKAMP, *supra* note 1, at 202 ("In some ways, the [1998 Copyright Term Extension Act] is the most blatant example of special-interest influence over the Copyright Act because retroactive term extensions for existing works do virtually nothing to promote innovation but they significantly burden future use, innovation, and expression by others.").

88. *Golan*, 132 S. Ct. at 888 ("The creation of at least one new work . . . is not the sole way Congress may promote knowledge and learning.").

89. *Id.*

90. *Id.* at 884.

But at this point, I might have overstepped into asking Bohannan and Hovenkamp to write a substantially different book. The literature on statutory interpretation is vast.[91] Bohannan and Hovenkamp have taken on so much so explicitly that they can easily be forgiven for declining to take substantial part in this fray as well. Bohannan and Hovenkamp have written a stimulating, rich, and instructive book. I do not agree with all of its contents, but I do believe strongly that readers will learn much from it. I encourage you to read Bohannan and Hovenkamp's book—and do so without restraint.

91. *See, e.g.*, WILLIAM N. ESKRIDGE, JR., PHILIP P. FRICKEY & ELIZABETH GARRETT, LEGISLATION AND STATUTORY INTERPRETATION 1 (2d ed. 2006) (observing that since "the mid-1970s" "a flood of scholarly and pedagogical materials on the legislative process and its products has inundated the law schools"); ABNER J. MIKVA & ERIC LANE, AN INTRODUCTION TO STATUTORY INTERPRETATION AND THE LEGISLATIVE PROCESS 51 (1997) ("The recent decade has seen an explosion in scholarly attention to statutory interpretation."); PETER L. STRAUSS, LEGISLATION: UNDERSTANDING AND USING STATUTES 396 (2006) (speaking of "the burgeoning literature about the new textualism and the problems of statutory interpretation").

Taking Hearers Seriously

BRANDISHING THE FIRST AMENDMENT: COMMERCIAL EXPRESSION IN AMERICA. By Tamara R. Piety. Ann Arbor, Michigan: University of Michigan Press, 2012. 342 pages. $70.00.

Reviewed by Burt Neuborne*

Introduction[1]

Once upon a time, vigorous Supreme Court enforcement of an expansive First Amendment was the darling of the American left. For most of the twentieth century, when progressive reformers were certain that they were on the right side of history, the left viewed free speech as a destabilizing force capable of eroding an oppressive and unequal status quo.[2] The possibility of negative fallout generated by an extremely robust First Amendment was deemed by people like me to be a small price to pay for the ability to invoke a robust free speech principle to usher in a better, more equal world.

---

* Inez Milholland Professor of Civil Liberties, New York University Law School. In the interests of full disclosure, in my capacities as a private lawyer, a staff lawyer for the American Civil Liberties Union for eleven years (I served as National Legal Director from 1981–1986), and as founding Legal Director of the Brennan Center for Justice at NYU Law School since the mid-1990s, I have participated as an advocate in many of the cases discussed in this Book Review. I make no claims to Olympian neutrality. If I did, no one would believe me.

1. The observations in this introductory material are not based on empirical data. Rather, they reflect my experience as a civil liberties lawyer for almost a half century and the evolution of my own views of the First Amendment. I use the categories of "left" and "right" loosely to reflect one's view of the existing economic and social structure. In my world, leftists tend to oppose the economic and social status quo as insufficiently egalitarian and unduly hierarchical. Those on the right tend to be more wedded to the economic and social status quo, either because they favor it, or are afraid that change will lead to something worse. Note, I confine my observations to the social and economic status quo. The political world does not lend itself to such generalizations. Characterizing the nature of one's approach to political change is uniquely dependent on the existing political baseline.

2. For example, the founding of the American Civil Liberties Union in 1919–1920 was largely driven by concern by leftists over the imprisonment and violent repression of labor organizers for the International Workers of the World (the IWW) and by the harsh treatment meted out to conscientious and political opponents to World War I. SAMUEL WALKER, IN DEFENSE OF AMERICAN LIBERTIES: A HISTORY OF THE ACLU 25–30 (1990). While the early ACLU defended Henry Ford's right to distribute an anti-Semitic newspaper, the bulk of its early caseload involved efforts to suppress speech critical of the status quo. *Id.* at 62, 68. For more on the history of the ACLU, see ROBERT C. COTTRELL, ROGER NASH BALDWIN AND THE AMERICAN CIVIL LIBERTIES UNION (2000), DIANE GAREY, DEFENDING EVERYBODY: A HISTORY OF THE AMERICAN CIVIL LIBERTIES UNION (1998), and WALKER, *supra*. For my modest contribution to ACLU history, see Burt Neuborne, *Of Pragmatism and Principle: A Second Look at the Expulsion of Elizabeth Gurley Flynn from the ACLU's Board of Directors*, 41 TULSA L. REV. 799 (2006).

Unlike the confident left, many mid-twentieth-century American conservatives, battered by the Great Depression of the 1930s, appalled by excesses committed in the name of conservative values by fascist lunatics,[3] and confronted by an almost unbroken phalanx of intellectual support for leftist programs, did not look to the future with confidence. Instead of viewing the uncensored exchange of views as a path to inevitable political, economic, and social triumph, many American conservatives viewed uncensored speech as a dangerous invitation to lawlessness and anarchy. The American right's unhappy role in the 1940s and early 1950s in connection with McCarthyism and the successful effort to outlaw the American Communist Party—ranging from enthusiastic leadership and support to tepid acquiescence—illustrates the fear of many mid-twentieth-century conservatives that uncensored speech and uncontrolled freedom of political association posed an unacceptable risk of radical social and economic change.[4]

When, in the late 1960s, the Warren Court protected the free speech rights of the Ku Klux Klan in *Brandenburg v. Ohio*,[5] formally rejecting the "bad tendency" test and transforming the Holmes/Brandeis dissents[6] into powerful legal doctrine highly protective of controversial speech, the left breathed a sigh of relief and awaited its inevitable triumph. The right hunkered down and vowed to fight on the beaches.[7] But a couple of

3. We now realize, of course, that right-wing lunatics had no monopoly on lethally oppressive behavior. The lunatic left more than held its own in that department, as can been seen, for example, in ALEKSANDR I. SOLZHENITSYN, THE GULAG ARCHIPELAGO 1918–1956: AN EXPERIMENT IN LITERARY INVESTIGATION (Thomas P. Whitney & Harry Willetts trans., Edward E. Ericson ed., 1978), and PHILIP SHORT, POL POT: ANATOMY OF A NIGHTMARE (2004).

4. *See generally* Dennis v. United States, 341 U.S. 494 (1951) (upholding criminal convictions of Communist Party leaders); ELLEN SCHRECKER, THE AGE OF MCCARTHYISM: A BRIEF HISTORY WITH DOCUMENTS (2d ed. 2002) (chronicling the curtailment of free speech and other civil liberties during the McCarthy era); ELLEN SCHRECKER, MANY ARE THE CRIMES: MCCARTHYISM IN AMERICA (1998) (same). For the legal history of the period, see GEOFFREY R. STONE, PERILOUS TIMES: FREE SPEECH IN WARTIME FROM THE SEDITION ACT OF 1798 TO THE WAR ON TERROR (2004).

5. 395 U.S. 444 (1969) (per curiam).

6. *Id.* at 447–48 & n.2. The two most celebrated Holmes/Brandeis First Amendment opinions are in *Whitney v. California*, 274 U.S. 357, 372 (1927) (Brandeis and Holmes, JJ., concurring), and *Abrams v. United States*, 250 U.S. 616, 624 (1919) (Holmes and Brandeis, JJ., dissenting). As we shall see, the two pioneering dissents reflect the two principal modern intellectual defenses of the free speech principle, with Brandeis stressing the dignitary interests of speakers, and Holmes stressing the instrumental value of free speech.

7. WILLIAM F. BUCKLEY, JR., GOD AND MAN AT YALE: THE SUPERSTITIONS OF "ACADEMIC FREEDOM" (1951), and BARRY GOLDWATER, THE CONSCIENCE OF A CONSERVATIVE (1960), are examples of defiant rejection of what appeared to conservatives to be liberal orthodoxy. The intellectual core of modern American conservatism was Russell Kirk's Ph.D. thesis entitled *The Conservative Mind from Burke to Santayana*, initially published in 1953. RUSSELL KIRK, THE CONSERVATIVE MIND FROM BURKE TO SANTAYANA (1953). The work has gone through multiple editions, with its title changed to *The Conservative Mind from Burke to Eliot*. RUSSELL KIRK, THE CONSERVATIVE MIND FROM BURKE TO ELIOT (7th rev. ed. 2001).

unexpected things happened on the First Amendment road to progressive paradise.

First, during the last two decades of the twentieth century, the intellectual core of the left's political agenda imploded, while the right enjoyed a remarkable intellectual renaissance.[8] Once the Berlin Wall fell in 1989, the left's political platform, premised on varying degrees of reliance on governmental redistribution of wealth—ranging from Marxism; to European democratic socialism; to the mild egalitarianism of the Kennedys and Lyndon Johnson's "War on Poverty"—ran headlong into an increasing sense that government—even democratic government—performs poorly as the economic or social linchpin of a society. Whether it was the grey tyranny of communism, the horrors of fascist rule, the kleptocratic antics of authoritarian dictators, or the often disheartening bureaucratic ineffectiveness of well-meaning welfare states, many—including many on the left—lost faith in the efficacy and moral legitimacy of a political agenda based on a strong, redistributive government. A generation of conservative intellectuals stepped into the programmatic vacuum, worshiping the market, glorifying individual autonomy, and questioning the role, indeed the very legitimacy, of much government regulation.[9] Not surprisingly, many on the left, faced with a newly confident right churning out ideas at a frantic pace, and lacking a coherent alternative political model of their own,[10] lost confidence in the inevitability of progressive change. Much leftist programmatic political speech dried up. What survived was a determined—and altogether noble—commitment to eliminating long-entrenched legal and social barriers to equal participation in the society. Since the achievement of such a political agenda actually weakens government by forbidding it from acting in certain discriminatory ways, and almost never asks more of government than negative prohibitions on categories of private discriminatory behavior that are already unpopular enough to be banned by the political majority, the American left continued to deploy a powerful rhetoric of formal equality, even as more ambitious speech about how to achieve real equality

8. For a description of the conservative intellectual renaissance, see GEORGE H. NASH, THE CONSERVATIVE INTELLECTUAL MOVEMENT IN AMERICA SINCE 1945 (2d ed. 2006), and JEFFREY HART, THE MAKING OF THE AMERICAN CONSERVATIVE MIND: *NATIONAL REVIEW* AND ITS TIMES (2005).

9. For an example of such conservatives, see ROBERT NOZICK, ANARCHY, STATE, AND UTOPIA (1974).

10. In fairness, John Rawls and Ronald Dworkin did all they could to generate an egalitarian intellectual position in JOHN RAWLS, A THEORY OF JUSTICE (1971), and RONALD DWORKIN, TAKING RIGHTS SERIOUSLY (1978). In the 1980s, Michael Walzer's SPHERES OF JUSTICE (1983) and Michael Sandel's work in the 1990s, LIBERALISM AND THE LIMITS OF JUSTICE (1998) and DEMOCRACY'S DISCONTENT: AMERICA IN SEARCH OF A PUBLIC PHILOSOPHY (1998), sought to provide intellectual alternatives to the free market. But they all relied upon a muscular redistributionist state to enforce egalitarian principles derived from the "veil of ignorance" or the best impulses of a society.

disappeared.[11] Speaking for myself, while such egalitarian rhetoric is admirable and important, I find it hard to convince myself that speech aimed at advancing a program of formal legal (as opposed to substantive economic and social) equality is so crucial to human progress that it justifies virtually any negative fallout from an extremely powerful First Amendment. That is a very different First Amendment cost–benefit ratio than the one I perceived as a young ACLU lawyer in the 1960s.

Second, flush with confidence and new ideas, the right discovered the First Amendment. During the 1970s, an expansive, judicially enforceable conception of free speech became as attractive to many on the right as it had historically been to the reformist left, ushering in an "era of First Amendment good feelings"[12] about the importance of an extremely strong First Amendment. The most dramatic manifestation of the "era of First Amendment good feelings" were the flag-burning cases in 1989 and 1990, when the right's newly minted dedication to an expansive First Amendment joined with the left's long-time commitment to expansive free speech to generate iconic 5–4 majorities (consisting of Justices Brennan, Marshall, Blackmun, Kennedy, and Scalia) upholding flag burning as protected speech.[13] In fact, the left–right First Amendment partnership had begun at least fifteen years earlier when Justice Harlan, a cautious conservative, provided the crucial fifth vote in *Cohen v. California*,[14] upholding the right to wear a jacket with the words "Fuck the Draft" emblazoned on the back.[15] The left and right deepened their First Amendment partnership in the mid-1970s, combining to hold in *Buckley v. Valeo* that the First Amendment protected the power of the super-rich to spend unlimited amounts of money

11. The unhappy fate of "affirmative action" in most legal contexts illustrates the limits of the formal egalitarian program. *See generally* Gratz v. Bollinger, 539 U.S. 244 (2003) (invalidating an affirmative action plan for university admissions); Adarand Constructors, Inc. v. Pena, 515 U.S. 200 (1995) (invalidating affirmative action in granting public construction contracts); Shaw v. Reno, 509 U.S. 630 (1993) (invalidating race-conscious reapportionment designed to increase minority representation); City of Richmond v. Croson, 488 U.S. 469 (1989) (invalidating racial set-asides for public construction projects). *But see generally* Grutter v. Bollinger, 539 U.S. 306 (2003) (upholding use of race as one criteria in law school admission). *Grutter* may be on borrowed time. The Supreme Court is scheduled to consider a similar issue in the 2012 term. Fisher v. Univ. of Tex., 132 S. Ct. 1536 (granting a petition for writ of certiorari to review the constitutionality of a university affirmative action program).

12. I am, of course, referring to the short period of virtually unanimous political support for President James Monroe from 1816–1820 discussed in GEORGE DANGERFIELD, THE ERA OF GOOD FEELINGS (1952).

13. Texas v. Johnson, 491 U.S. 397, 398, 406 (1989); United States v. Eichman, 496 U.S. 310, 311, 319 (1990). Justice Brennan wrote for the Court in both cases. *Johnson*, 491 U.S. at 398; *Eichman*, 496 U.S. at 311. The dissenters were Chief Justice Rehnquist and Justices White, Stevens, and O'Connor. *Johnson*, 491 U.S. at 421, 436; *Eichman*, 496 U.S. at 319.

14. 403 U.S. 15 (1971).

15. Justice Harlan was joined by Justices Douglas, Marshall, Brennan, and Stewart. *Cohen*, 403 U.S. at 15. Chief Justice Burger, joined by Justices Black, White, and Blackmun, dissented. *Cohen*, 403 U.S. at 27, 28.

to affect electoral outcomes.[16] *Buckley* gave the 1% a tangible reason to celebrate a muscular First Amendment. *Buckley* was closely followed by *Virginia State Board of Pharmacy v. Virginia Citizens Consumer Council*, in which a coalition of liberals and conservatives overturned *Valentine v. Chrestensen*[17] and recognized limited but important First Amendment protection for truthful, nonmisleading commercial advertising.[18] *Virginia Pharmacy* gave corporate management a strong stake in the First Amendment. In the late 1970s, the Court's liberals and conservatives joined once again in *First National Bank of Boston v. Bellotti*[19] to recognize a corporate free speech right to use corporate treasury funds to oppose a referendum on raising taxes.[20] *Bellotti* raised corporate America's already substantial stake in free speech even higher. In the 1980s, the Court's left and right wings joined to recognize the First Amendment as a potent shield against government efforts to regulate massive concentrations of communicative power, endearing the First Amendment to Rupert Murdoch and his friends.[21] To add insult to injury (literally), in the 1990s, the Court

16. 424 U.S. 1, 58 (1976). The fragmented series of per curiam and individual opinions in *Buckley* usually shakes out to 7–1, with Chief Justice Burger dissenting and Justice Stevens not participating.

17. 316 U.S. 52 (1942).

18. 425 U.S. 748, 773 (1976). Justice Blackmun wrote for seven Justices, including Brennan and Marshall. *Id.* at 749. Chief Justice Burger wrote a concurrence. *Id.* at 773. Justice Rehnquist was the lone dissenter. *Id.* at 781. Justice Stevens did not participate. *Id.* at 773. The commercial speech doctrine received its fullest articulation several years later in *Cent. Hudson Gas & Elec. Corp. v. Public Serv. Comm'n of N.Y.*, 447 U.S. 557 (1980). The eight Justices who voted to invalidate a ban on promotional messages by electric companies found it very difficult to identify exactly what falls under commercial speech, proffering four different tests. *Id.* at 564. Chief Justice Rehnquist continued to dissent from the grant of broad First Amendment power to corporations. *Id.* at 583.

19. 435 U.S. 765 (1978).

20. *Id.* at 784–86, 795.

21. Although the Court rejected a right to reply to press attacks in *Miami Herald Publ'g Co. v. Tornillo*, 418 U.S. 241, 258 (1974), the Court was initially receptive to government efforts to provide dissenting voices with access to the broadcast media. *See, e.g.*, FCC v. Nat'l Citizens Comm. for Broad., 436 U.S. 775, 779 (1978) (upholding the ban on cross-ownership of a newspaper and a TV station in same market); Red Lion Broad. Co. v. FCC, 395 U.S. 367, 375 (1969) (upholding the "fairness doctrine"); Associated Press v. United States, 326 U.S. 1, 12–20 (1945) (applying antitrust laws to media setting). After the FCC rejected a fairness doctrine challenge to a broadcaster's policy of refusing to accept paid editorial advertisements, the Court declined to recognize a First Amendment right of access to broadcast media. Columbia Broad. Sys. v. Democratic Nat'l Comm., 412 U.S. 94, 121–32 (1973). The autonomy of broadcasters was upheld in several subsequent cases. *See* Ark. Educ. Television Comm'n v. Forbes, 523 U.S. 666, 683 (1998) (upholding a broadcaster's exclusion of a candidate from debate on public TV); Turner Broad. Sys., Inc. v. FCC (*Turner I*), 512 U.S. 622, 637 (1994) (rejecting application of *Red Lion* to cable broadcasting); FCC v. League of Women Voters of Cal., 468 U.S. 364, 402 (1984) (invalidating a ban on editorials by a public TV stations). The chaotic state of current law on media diversity is reflected in *Prometheus Radio Project v. FCC*, 373 F.3d 372, 382 (3d Cir. 2004), *Sinclair Broad. Grp., Inc. v. FCC*, 284 F.3d 148, 152 (D.C. Cir. 2002), and *Fox Television Stations, Inc. v. FCC*, 280 F.3d 1027, 1033 (D.C. Cir. 2002), each of which considers an FCC rule on media ownership.

invoked the First Amendment to strike down bans on hate speech targeting vulnerable minorities.[22]

It did not take long for some on the left to suspect that they had entered into an unequal First Amendment partnership.[23] By 1990, some progressive reformers feared that an expansive First Amendment doctrine that allows scruffy kids to burn flags and provides tepid protection for street demonstrations[24] also protects (1) massive commercial spending by corporations and individuals aimed at selling their products or polishing their images; (2) uncontrolled corporate speech aimed at shaping public opinion on political, economic, and social issues; (3) uncontrolled campaign spending by the super-rich—including corporations—designed to affect electoral outcomes; (4) concentration of media power in a handful of huge corporations that own or control virtually all of the nation's newspapers, television and radio stations, and book publishers; and (5) bursts of verbal venom aimed at historically weak targets seeking access to education and decent housing—hardly a prescription for progressive change. They began to view such a one-sided First Amendment partnership as a Faustian bargain, far more likely to reinforce the status quo than to destabilize it in pursuit of greater equality and less hierarchy.

By 2000, the era of First Amendment good feelings was over, a victim of the deregulatory impact of an extremely robust First Amendment. The right rejected aspects of First Amendment deregulation, rediscovering the

---

22. *See, e.g.*, Virginia v. Black, 538 U.S. 343, 348, 367–68 (2002) (invalidating a conviction for cross burning, a specialty of the Ku Klux Klan, in the absence of specific proof of intent to intimidate); R.A.V. v. City of St. Paul, 505 U.S. 377, 381 (1992) (declaring unconstitutional a city ordinance criminalizing cross burning on the grounds that the ordinance "prohibits otherwise permitted speech solely on the basis of the subjects the speech addresses").

23. For early expressions of concern, see CASS R. SUNSTEIN, DEMOCRACY AND THE PROBLEM OF FREE SPEECH (1992). Despite my admiration for Sunstein's work generally, I wrote a skeptical review of DEMOCRACY AND THE PROBLEM OF THE FREE SPEECH that parallels this essay in many ways in Burt Neuborne, *Blues for the Left Hand: A Critique of Cass Sunstein's* Democracy and the Problem of Free Speech, 62 U. CHI. L. REV. 423 (1995).

24. The Supreme Court's cautious approach to what some call "body rhetoric," e.g., Kristine M. Zaleskas, *Pride, Prejudice or Political Correctness? An Analysis of* Hurley v. Irish-American Gay, Lesbian & Bisexual Group of Boston, 29 COLUM. J.L. & SOC. PROBS. 507, 533 (1996)—picketing, marching, and camping out in public, usually by relatively poor protestors—is illustrated by *Clark v. Cmty. for Creative Non-Violence*, 468 U.S. 288 (1984), in which the Court upheld a National Park Service regulation barring camping in certain parks as applied to prevent activists from sleeping in two Washington, D.C. parks as part of a demonstration against homelessness. *Id.* at 289. Occupy Wall Street was unable to establish semi-permanent encampments in parks and other public spaces, and experienced repeated difficulty with police in mounting mass marches. No, Virginia, the streets do not belong to the people, at least not the poor people. *See, e.g.*, James Barron & Colin Moynihan, *City Reopens Park After Protesters Are Evicted*, N.Y. TIMES, Nov. 15, 2011, *available at* http://www.nytimes.com/2011/11/16/nyregion/police-begin-clearing-zuccotti-park-of-protesters.html (noting that a "judge upheld the city's move to clear the park and bar the protesters from bringing back their tents or sleeping overnight"); *see also* Ashutosh A. Bhagwat, *Assembly Resurrected*, 91 TEXAS L. REV. 351 (2012) (reviewing JOHN D. INAZU, LIBERTY'S REFUGE: THE FORGOTTEN FREEDOM OF ASSEMBLY (2012)); Timothy Zick, *Recovering the Assembly Clause*, 91 TEXAS L. REV. 376 (2012) (same).

attractions of hierarchy and control by shrinking public employee free speech rights,[25] limiting the free speech rights of students,[26] and blocking the ability of Americans to interact peacefully with foreign organizations labeled by the government as "terrorist."[27] The left realized the danger to egalitarian values posed by a strongly deregulatory First Amendment, opposing First Amendment decisions that upped the right's free speech ante even further by construing the First Amendment as protecting uncontrollable electoral spending by corporations,[28] while simultaneously invoking the First Amendment to restrict political spending by unions and to invalidate efforts to use matching funds as a practical method of publicly funding political campaigns.[29]

Faced with the emergence of potent First Amendment doctrine that appears to some to favor the rich and powerful, progressives reacted in three ways. Some, like the ACLU and a respected cadre of lawyers and academics led by Floyd Abrams and Kathleen Sullivan, argue that protecting commercial speech, corporate political speech, hate speech, and the uncontrolled electoral spending power of the super-rich is simply the necessary and logical consequence of vigorous enforcement of a robust free speech principle.[30] Progressives, they argue, should celebrate such a potent

---

25. *See, e.g.*, Garcetti v. Ceballos, 547 U.S. 410, 420–26 (2006) (upholding the dismissal of a deputy district attorney for internal criticism of a failure to respond to misrepresentations in a search warrant); Waters v. Churchill, 511 U.S. 661, 679–82 (1994) (holding that the statements of a nurse criticizing hospital operations were constitutionally unprotected due to their "disruptive" nature, such that proof that the nurse's discharge was based entirely on these statements would defeat the nurse's civil rights claim against hospital); Connick v. Myers, 461 U.S. 138, 154 (1983) (limiting employee freedom to circulate internal criticism of an employer).

26. *See, e.g.*, Morse v. Frederick, 551 U.S. 393, 396–97 (2007) (upholding discipline for a student who displayed a banner with drug connotations a at school function); Hazelwood Sch. Dist. v. Kuhlmeier, 484 U.S. 260, 273, 276 (1988) (upholding a principal's editorial control over the official student newspaper); Bethel Sch. Dist. No. 403 v. Fraser, 478 U.S. 675, 681–86 (1986) (upholding discipline for a student who delivered a school-office nomination speech containing sexual innuendos).

27. *See, e.g.*, Holder v. Humanitarian Law Project, 130 S. Ct. 2705, 2722–31 (2010) (denying First Amendment speech and assembly challenges to a federal statute criminalizing the provision of "material support or resources" to foreign organizations determined to engage in terrorist activity).

28. *E.g.*, Citizens United v. FEC, 130 S. Ct. 876, 929 (2010) (Stevens, J., concurring in part and dissenting in part).

29. *E.g.*, Knox v. Serv. Emps. Int'l Union, 132 S. Ct. 2277, 2290–91, 2293 (2012) (suggesting in dicta joined by four Justices that public employees should be required to "opt in" to support their unions in using mandatory dues for political purposes); Ariz. Free Enter. Club's Freedom Club PAC v. Bennett, 131 S. Ct. 2806, 2812–13 (2011) (invalidating an Arizona matching-fund law by a 5–4 vote).

30. *See* FLOYD ABRAMS, SPEAKING FREELY: TRIALS OF THE FIRST AMENDMENT 234–35 (2006) (noting that Abrams, Sullivan, and the ACLU believe the First Amendment prohibits limits on expenditures for political campaigns); *id.* at 245 (urging liberals to resist limitations on commercial speech); Floyd Abrams, *Hate Speech: The Present Implications of a Historical Dilemma*, 37 VILL. L. REV. 743, 755–56 (1992) (arguing that the First Amendment demands the protection of hate speech); Kathleen M. Sullivan, *Free Speech Wars*, 48 SMU L. REV. 203, 204, 214 (1994) (opposing regulation of hate speech on First Amendment grounds); *ACLU History: Taking a Stand for Free Speech in Skokie*, ACLU (Sept. 1, 2010), http://www.aclu.org/free-

First Amendment as a great shield of freedom, and get their acts together to compete intellectually in such a *laissez faire* marketplace of ideas.[31] According to the ACLU, resource imbalance between or among competing speakers should be dealt with not by censoring the strong speakers, but by subsidizing the weak ones.[32]

Others, like me, continue to pledge allegiance to a vigorous First Amendment in most settings, but argue that the Court's reasoning in cases like *Buckley v. Valeo*, *Citizens United*, and *Arizona Free Enterprise* confuses "conduct" in the form of massive spending with "speech" and undervalues the governmental interests in maintaining electoral equality and avoiding the appearance of systemic political corruption. Faced with an apparently implacable five-Justice majority supporting *Buckley* and *Citizens United*, however, such an incremental reformist approach appears to some to border on the quixotic.[33]

Beginning in the early 1990s, a third group of progressive intellectuals, troubled by the Faustian bargain, unwilling to accept the ACLU's unyielding iron vision of the First Amendment, and impatient with a moderately reformist legal strategy that did not appear to promise short-term results, challenged not merely the reasoning of the Court's five-Justice majority in cases like *Buckley* and *Citizens United*, but the very notion that regulating speech is particularly antithetical to a free society.[34] They go for the First Amendment jugular by challenging the underlying intellectual justifications for treating free speech as a trumping value that overrides almost all good faith, plausible efforts at government regulation. Professor Tamara R. Piety's passionately written and disturbing book, *Brandishing the First*

---

speech/aclu-history-taking-stand-free-speech-skokie (explaining the ACLU's free-speech defense of a neo-Nazi group banned from marching through Skokie, Illinois).

31. *See* ABRAMS, *supra* note 30, at 232 (warning that citizens should be wary of any regulation of speech, especially bans on campaign ads); Sullivan, *supra* note 30, at 213 (advocating deregulation of the marketplace of ideas); *The ACLU and* Citizens United, ACLU (Mar. 27, 2012), http://www.aclu.org/free-speech/aclu-and-citizens-united ("Our system of free expression is built on the premise that the people get to decide what speech they want to hear; it is not the role of the government to make that decision for them.").

32. *The ACLU and* Citizens United, *supra* note 31.

33. In *American Tradition Partnership v. Bullock*, 132 S. Ct. 2490 (2012), the Court declined Montana's request to reconsider the dicta in *Citizens United* recognizing the First Amendment right of multi-shareholder, for-profit corporations to spend unlimited sums to affect the outcome of an election. *Id.* at 2491 (per curiam). The four liberal members of the Court expressed a willingness to hear the case, but declined to invoke the "rule of four" to place the case on the plenary docket because they perceived no softening in the position of the five-Justice conservative majority in *Citizens United*. *Id.* at 2491–92 (Breyer, J., dissenting). Frankly, that's why liberals lose so often. I do not believe that four conservative Justices would have passed on the chance to force a public reconsideration of a case like *Citizens United*.

34. *See, e.g.*, Reza R. Dibadj, *The Political Economy of Commercial Speech*, 58 S.C. L. REV. 913, 915 (2007) (arguing against the use of the First Amendment as a way to "sidestep economic regulation" and "grant ever-expansive rights to commercial speech"); Sylvia A. Law, *Addiction, Autonomy, and Advertising*, 77 IOWA L. REV. 909, 912 (1992) (offering a First Amendment analysis which "permits significant legal constraint" on certain types of commercial speech).

Amendment,[35] is the latest entry in the left's skeptical reexamination of much of current First Amendment doctrine. Professor Piety pursues three general goals: (1) confining commercial speech protection to truthful, nonmisleading speech; (2) defining commercial speech very expansively in an effort to subject speech by corporations to greater government regulation; and (3) assaulting the very idea of constitutionally protected corporate speech. Much of her book is a moving *cri de coeur* about the hijacking of the First Amendment by large for-profit corporations and the super-rich, resulting in an overwhelming outpouring of privately funded speech by rich and powerful speakers designed to manipulate the population into: (1) buying products (often unneeded) that are essentially identical to competing products in everything but manipulative advertising; (2) accepting slick, but one-sided, corporate-funded arguments about important public policy issues that are driven solely by the short-term profit interests of corporate management; and (3) acquiescing in the continued political stranglehold of the super-rich on American democracy.

Although Professor Piety confines her critique to commercial and corporate speech, many of her arguments cut to the bone of much traditional free speech protection. She begins by reminding us that the emergence of free speech as a trumping constitutional value is a relatively recent phenomenon in American law. For most of the nation's history, she notes, free speech was simply one of a number of important values to be balanced and blended in the formation of American democracy.[36] Although she does not do so, she could have noted, as well, that no other functioning democracy espouses our current willingness to allow free speech values to trump virtually all efforts at government regulation of the communicative process, even when the regulation seems well-intentioned and is plausibly justified.[37] Professor Piety's critique of commercial and corporate speech then turns to the three sets of reasons—I call them "dignitary," "instrumental," and "cautionary"—that are usually invoked to explain why we treat speech taking place in Mr. Madison's neighborhood[38] so differently from other forms of potentially regulable behavior.

Five kinds of people live in Mr. Madison's First Amendment neighborhood—speakers, hearers, conduits (who transmit the speech of

35. TAMARA R. PIETY, BRANDISHING THE FIRST AMENDMENT: COMMERCIAL EXPRESSION IN AMERICA (2012).

36. *See* PIETY, *supra* note 35, at 17, 55–56 (2012) (positing that the early twentieth century marks a break with prior First Amendment law).

37. For a useful survey of the treatment of free speech by our sister democracies, see Adrienne Stone, *The Comparative Constitutional Law of Freedom of Expression* (Melbourne Law Sch., Legal Studies Research Paper No. 476, July 1, 2010), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1633231.

38. All free speech under the First Amendment occurs in the neighborhood that James Madison built. He introduced what became the First Amendment on the floor of Congress and shepherded the Amendment through to ratification. James H. Read, *James Madison*, *in* 2 ENCYCLOPEDIA OF THE FIRST AMENDMENT 699, 699–701 (John R. Vile et al. eds., 2009).

others), government regulators (you can tell them by their shifty eyes and black hats), and speech targets (the subjects of the speech in question). Since the mass media has persuaded the Supreme Court that conduits should usually be treated as speakers,[39] we can ignore them for the purposes of this essay. We can also ignore the speech targets. They are the neighborhood slum dwellers, whose interests are almost always subordinated to those of speakers and hearers.[40] Since the decision to subordinate the interests of speech targets to the interests of speakers, hearers, and conduits is beyond the scope of this essay,[41] that leaves speakers, hearers, and regulators.

Speakers are the neighborhood aristocrats. Most First Amendment theory and doctrine is unabashedly speaker centered.[42] Hearers are the neighborhood *haute bourgeoisie*—privileged and influential, but subordinate to speakers. When the interests of speakers and hearers differ, the edge usually goes to speakers.[43] Hearers may, however, assert an independent

39. *See generally* Turner Broad. Co. v. FCC, 512 U.S. 622, 643–44 (1994) (speaking of "must-carry" regulation of cable providers as being regulation of their "speech"); Miami Herald Publ'g Co. v. Tornillo, 418 U.S. 241, 243, 258 (1974) (holding that a newspaper is more than a "passive receptacle or conduit" and that the First Amendment protects an editor's discretion from a law requiring equal page space for a criticized political candidate); Columbia Broad. Sys., Inc. v. Democratic Nat'l Comm., 412 U.S. 94, 127, 132 (1973) (rejecting a right of access to purchase editorial air time on broadcast media and acknowledging the First Amendment protections due to broadcast licensees).

40. *See, e.g.*, Snyder v. Phelps, 131 S. Ct. 1207, 1219 (2011) (upholding the First Amendment right to engage in deeply offensive picketing in the vicinity of a fallen soldier's funeral); Hustler Magazine, Inc. v. Falwell, 485 U.S. 46, 56–57 (1988) (reaffirming the protected nature of intentionally hurtful speech); Smith v. Collin, 439 U.S. 916, 916 (1978) (Blackmun and White, JJ., dissenting from denial of cert.) (declining to review decision protecting the right of Nazis to march through a village inhabited by thousands of Holocaust survivors); Nat'l Socialist Party v. Village of Skokie, 432 U.S. 43, 43–44 (1977) (staying a preliminary injunction of a proposed march by the same group); Cohen v. California, 403 U.S. 15, 26 (1971) (protecting speech deeply offensive to hearers); N.Y. Times Co. v. Sullivan, 376 U.S. 254, 292 (1964) (protecting false defamatory speech as long as the speaker believes it to be true).

41. The high water mark of concern for speech targets occurred in *Beauharnais v. Illinois*, 343 U.S. 250, 258–61 (1952) (upholding criminal prosecutions under group libel laws). Jeremy Waldron has published a deeply felt plea for better treatment of speech targets. JEREMY WALDRON, THE HARM IN HATE SPEECH (2012).

42. *See, e.g.*, United States v. Alvarez, 132 S. Ct. 2537, 2548 (2012) (invalidating a federal statue criminalizing false claims of having received military decorations); *Snyder*, 131 S. Ct. at 1213, 1219 (upholding the right of virulent antigay activists to demonstrate in the close vicinity of the funeral of a fallen soldier); Texas v. Johnson, 491 U.S. 397, 418–19 (1989) (upholding flag burning despite its deeply offensive nature); Hess v. Indiana, 414 U.S. 105, 107–09 (1973) (categorizing "[w]e'll take the fucking street later" as protected speech); Cohen v. California, 403 U.S. 15, 16–17 (1971) (upholding a man's right to wear a jacket reading "Fuck the Draft" in a municipal courthouse). The hearers' interest may, however, prevail in a few settings where it approaches an independent constitutional value. *See, e.g.*, Hill v. Colorado, 530 U.S. 703, 725–35 (2000) (upholding ordinance precluding antiabortion protestors from approaching to within eight feet of an unwilling patron of abortion facility); Frisby v. Schultz, 487 U.S. 474, 483 (1988) (upholding a ban on targeted picketing of individual residences but permitting roving picketing).

43. *See, e.g.*, R.A.V. v. City of St. Paul, 505 U.S. 377, 380, 391 (1992) (holding a statute banning cross burning facially unconstitutional); Virginia v. Black, 538 U.S. 343, 347–48 (2003) (same).

right to receive information that is not dependent on the existence of a protected speaker.[44] As *Virginia Pharmacy* and *Bellotti* demonstrate, that is where commercial and corporate speech ultimately comes from.[45] As a practical matter, such a hearer-centered First Amendment right has generally been applied by the Court to benefit an otherwise unprotected speaker—even, as Professor Piety points out, when many hearers don't want to hear the speech.[46]

Finally, government regulators are treated like the neighborhood motorcycle gang, prone to terrorizing the residents unless carefully policed.[47]

Speakers and hearers defend their privileged status by invoking their dignity as autonomous human beings blessed with free will. Defenders of a robust free speech principle almost always begin by arguing that autonomous human beings must be able to speak and hear freely in order to shape their own destinies and form their own preferences.[48] In commercial and corporate speech settings, where speakers appear to lack the necessary human dignity, unprotected speakers are often permitted to rely on the rights of autonomous hearers, whose dignity is said to require uncensored access to the speech in question.[49]

---

44. *See* Lamont v. Postmaster Gen., 381 U.S. 301, 305 (1965) (upholding First Amendment right to receive mailings from foreign governments that lack First Amendment rights). It is hard to believe, but *Lamont* was the first case striking down an act of Congress under the First Amendment. Memorandum from Laurence H. Tribe to Congress 20 (Dec. 6, 2011), *available at* http://www.scribd.com/doc/75153093/Tribe-Legis-Memo-on-SOPA-12-6-11-1.

45. First Nat'l Bank of Bos. v. Bellotti, 435 U.S. 765, 783 (1978); Va. State Bd. of Pharm. v. Va. Citizens Consumer Council, Inc., 425 U.S. 748, 756 (1976).

46. Citizens United v. Fed. Election Comm'n, 130 S. Ct. 876, 913 (2010) was decided in an era when polls indicate that the overwhelming majority of the population does not wish to be subjected to a corporate electoral barrage. *See* Susan Page, *Swing States Poll: Amid Barrage of Ads, Obama Has Edge*, USA TODAY, July 8, 2012, http://usatoday30.usatoday.com/news/politics/story/2012-07-08/swing-states-poll/56097052/1 (stating four out of five respondents in swing states subject to heavy campaign advertising could not wait for the election season to end).

47. Virtually every Supreme Court First Amendment decision, especially the prior-restraint, overbreadth, void-for-vaguenesss, and equality doctrines, contains language warning of the risks of vesting government with power to censor, especially when the power is poorly defined. *E.g.*, *Cohen*, 403 U.S. at 26; *Citizens United*, 130 S. Ct. at 907.

48. *See* Martin H. Redish, *The Value of Free Speech*, 130 U. PA. L. REV. 591, 607 (1982) (arguing that the individual needs a "free flow of information" and opinion related to "life-affecting decisions"); David A. Strauss, *Persuasion, Autonomy, and Freedom of Expression*, 91 COLUM. L. REV. 334, 357 n.64 (1991) (asserting that an essential idea of "autonomy is that there are 'capacities central to human rationality' that an autonomous person must be free to exercise"). The historic evolution of a dignitary explanation for a robust free speech principle runs from Immanuel Kant to John Locke to John Milton. *See generally* David A.J. Richards, *Free Speech and Obscenity Law: Toward a Moral Theory of the First Amendment*, 123 U. PA. L. REV. 45, 46 (1974) (arguing that "there is little question that the [First] [A]mendment was part of and gives expression to a developing moral theory regarding the equal liberties of men which had been given expression by Milton and Locke and which was being given or was to be given expression by Rousseau and Kant" (footnotes omitted)).

49. *See Citizens United*, 130 S. Ct. at 898–99 (noting, in a case involving corporate speech, that the government cannot, by inhibiting the flow of speech, deprive the public of the chance to evaluate the speech for themselves); *Bellotti*, 435 U.S. at 790–92 (same); *Va. State Bd. of Pharm.*,

Once discussion of the Kantian vision of an autonomous human being[50] is exhausted (it can take days), speakers and hearers generally retreat to a less ontological position. The free flow of ideas and information, they argue, is instrumentally essential to the functioning of crucial institutions like democracy, markets, and scientific inquiry.[51] At this point, the ghost of Galileo is usually trotted out to demonstrate both the affront to human dignity and the adverse impact on scientific inquiry imposed by the Church's censorship of his work.[52]

Finally, like the privileged of any neighborhood, speakers and hearers defend their status by fear-mongering, reminding us that government censors have historically behaved very badly, often using the state's monopoly of force to crush dissenters and to perpetuate the censor's grip on power. The only safe solution, argue First Amendment stalwarts, is a set of prophylactic First Amendment rules—both substantive and procedural—designed to prevent the motorcycle gang from getting any traction.[53]

I find the combined impact of the three arguments—dignitary, instrumental and cautionary—very convincing. Galileo always gets me. That is why I strongly prefer incremental doctrinal fixes for cases like *Buckley* and *Citizens United* to a frontal assault on the First Amendment's intellectual underpinnings. I concede, though, that as Professor Piety

---

425 U.S. at 763–65 (stating the importance of the free of flow of commercial information to consumers and applying the First Amendment outside of the context of public discourse).

50. *See* IMMANUEL KANT, THE METAPHYSICS OF MORALS 223 (Mary Gregor ed., Cambridge Univ. Press 2003) (stressing the fundamental autonomy of the individual).

51. *See* JOHN MILTON, AREOPAGITICA: A PLEA FOR UNLICENSED PRINTING 5–6 (J.W. Hales ed., Oxford, Clarendon Press, 1874) (1644); JOHN STUART MILL, ON LIBERTY 9–32 (London, Longmans, Green, Reader & Dyer 1880) (1859) (stressing the instrumental value of free speech in the search for truth). Oliver Wendell Holmes and Alexander Meiklejohn are leading modern instrumentalists as can be seen in *Abrams v. United States*, 250 U.S. 616, 616–17 (1919) (Holmes, J., dissenting), and Alexander Meiklejohn, *The First Amendment Is an Absolute*, 1961 Sup. Ct. Rev. 245, 256–57 (setting forth certain vital societal values which the First Amendment serves to protect).

52. *See* MAURICE A. FINNOCHIARIO, THE GALILEO AFFAIR: A DOCUMENTARY HISTORY 3 (1989) (asserting that Galileo's trial has been a constant reference point for later scientific critics of religion). Milton's *Areopagitica*, one of the landmarks in the evolution of free speech theory, was almost certainly influenced by the young Milton's visit to Galileo during Galileo's house arrest. *See* 1 WILLIAM RILEY PARKER, MILTON: A BIOGRAPHY 179 (Gordon Cambell ed., 2d ed.1996) (asserting that Milton's views on the "evils of censorship" were heavily influenced by Galileo).

53. Frederick Schauer is one of the leading proponents of the cautionary approach to government censorship. *See* FREDERICK SCHAUER, FREE SPEECH: A PHILOSOPHICAL ENQUIRY 136–45 (1982) (explaining how the uncertainty principle undergirds free speech doctrine). The cautionary approach, in its protectiveness of the right's exercise, leads to such prophylactic protections as the ban on prior restraints, the overbreadth and vagueness doctrines, and the insistence that like speakers be treated alike. *See id.* at 138–40 (stating the free speech principle requires some higher standard of proof, which necessarily leads to the failure to prevent some dangers in the interest of protecting the principle).

trenchantly argues, each of the three pillars underlying Mr. Madison's neighborhood displays large and unsightly cracks.[54]

First, Professor Piety argues that the dignitary concept of an "exogenous" rational human being, operating as an autonomous creature generating and processing information as the raw material for constructing her own personality and preferences, appears to be descriptively inaccurate.[55] She concedes that the concepts of free will and autonomous rationality are important as an ideal (indeed, I believe that they are a necessary existential fiction), but argues that they do not describe the real world, especially the real world of commercial and corporate speech.[56] Professor Piety argues persuasively that commercial sellers engaged in hawking their wares and for-profit corporations advancing their short-term profit interests are not dignitary speakers, and that much of what passes for advertising, marketing, and corporate public-relations speech contains little usable information.[57] Rather, she contends that it is a tissue of manipulative techniques designed to

54. Professor Piety organizes her critique of the intellectual justification for commercial and corporate speech by testing both against Tom Emerson's classic defense of the free speech principle as discussed in THOMAS I. EMERSON, THE SYSTEM OF FREE EXPRESSION (1970). PIETY, *supra* note 35, at 56–60); *see also* Thomas I. Emerson, *Towards a General Theory of the First Amendment*, 72 YALE L.J. 877, 878–79 (1963) (outlining a basic four-part defense of free speech). Emerson cites four reasons to protect free speech: respect for human dignity; the importance of free speech in the search for truth; the importance of democratic participation; and the need to draw a line between change and stability. *Id.* at 878–86. Emerson's first argument is obviously dignitary, while reasons two, three, and four are instrumental. While Emerson's work is one of the milestones in modern First Amendment thought, it is somewhat dated. As Professor Piety notes, much of Emerson's analysis is based on a bright-line distinction between speech and conduct that has not proven administrable. PIETY, *supra* note 35, at 55. Moreover, the Emerson formulation omits one important free speech argument—the special risks associated with government censorship. Finally, he fails to discuss either commercial or corporate speech, no doubt because he wrote prior to their recognition by the Court. Thus, while Emerson's work is important, it is not a foolproof barometer and seems a curious choice as the definitive repository of First Amendment wisdom. It does, however, set forth the dignitary and instrumental arguments for free speech in a way that allows Professor Piety to impose a coherent structure on her critique.

55. Professor Piety is, of course, not the first to question the existence of genuinely "exogenous" human rationality machines. *See generally* SUNSTEIN, *supra* note 23, at 137–45 (criticizing both autonomy and rationality as bases for free speech rights). Much of the best discussion of autonomy in a First Amendment context occurs in Richard H. Fallon, Jr., *Two Senses of Autonomy*, 46 STAN. L. REV. 875 (1994) (introducing a framework that includes descriptive and ascriptive autonomy and concluding that these concepts often pull in opposite directions thereby complicating First Amendment problems). *Compare also* Thomas Scanlon, *A Theory of Free Expression*, 1 PHIL. & PUB. AFF. 204, 217 (1972) ("The harm of coming to have false beliefs is not one that an autonomous man could allow the state to protect him against through restrictions on expression."), *with* T.M. Scanlon, Jr., *Freedom of Expression and Categories of Expression*, 40 U. PITT. L. REV. 519, 533–34 (1979) ("My argument for the Millian Principle . . . employed the idea of autonomy . . . as a constraint on justifications of authority . . . . The idea of such a constraint now seems to me mistaken."). Professor Piety adopts Fallon's terminology of "descriptive" and "ascriptive" autonomy, but not the subtlety with which he discusses it. PIETY, *supra* note 35, at 81–82. She does almost nothing with the ascriptive nature of the concept.

56. PIETY, *supra* note 35, at 86–87.

57. *See id.* at 88–106 (discussing the concept of brand and how advertisers develop that concept in consumers).

play on a hearer's emotions and nonrational needs.[58] She cites modern psychological research suggesting that human beings are neither freestanding nor autonomous, but are really malleable constructs of the information bath into which they are born and within which they spend their lives.[59]

Descriptively, Professor Piety may well be right about the limits of free will. After all, that is why we have compulsory public education, the primary purpose of which is to inculcate the young with prevailing community values. Thus, the image of a heroic First Amendment arming Prometheus to rebel against the gods—or, at least, General Electric—may well be an urban myth. If, argues Professor Piety, human beings are really malleable creatures whose personality and preferences are substantially shaped by the social, economic, and political soup in which they swim, why not admit that public education functions from the cradle to the grave and get on with the task of providing excellent government guidance?[60]

Professor Piety responds briefly to the cautionary argument about empowering the neighborhood motorcycle gang with power to censor by noting that if government is banished from the process of regulating commercial and corporate speech, the regulatory vacuum will inevitably be filled by private employers and self-interested private sources of guidance and control. Moreover, those are the very types of entities who often seek to manipulate preferences without a hearer's conscious knowledge by using techniques pioneered on Madison Avenue and in the totalitarian square.[61] She analogizes legal doctrine exposing vulnerable hearers to such a potentially harmful speech barrage as a misguided form of "tough love."[62]

Although Professor Piety confines her critique of the dignitary justification for free speech to commercial and corporate speech, she makes no effort to explain why it does not also erode the intellectual foundation for extensive protection of controversial political speech. Manipulative Madison Avenue techniques that appeal to the emotions as opposed to the rational mind and the skillful use of the "big lie" are not confined to the commercial or corporate sphere. They permeate our political discourse. In the end, I fear that a creature as weak and malleable as Professor Piety's condescending portrait of a typical hearer is a poor candidate for democratic self-governance.[63] Despite our descriptive shortcomings, it is, I believe,

---

58. *See id.* at 108–20 (examining various cognitive biases and explaining how marketers manipulate those biases).

59. *Id.* at 108–15.

60. *Id.* at 99–104.

61. *See id.* at 133 (noting that a lack of government intervention leaves society "at the mercy of professional persuaders").

62. *Id.* at 121–22.

63. As I read Professor Piety's description of a malleable hearer shaped by the forces of darkness, I couldn't help thinking of Herbert Marcuse's attack on the idea of the autonomous self in Herbert Marcuse, *Repressive Tolerance*, *in* ROBERT PAUL WOLFF ET AL., A CRITIQUE OF PURE TOLERANCE 81, 86–87 (1965).

existentially necessary for the law to treat us as autonomous creatures blessed with free will and the capacity for rational choice; first, because such a leap of legal faith reinforces the Kantian ideal towards which we should strive; second, because once you remove the crucial component of free will (even if it is a fiction) from the human equation, the argument for democratic governance unravels into the nightmare of "false consciousness";[64] and third, because, whatever our descriptive shortcomings, I believe that most human beings stubbornly demonstrate substantial, if not perfect, ability to think for themselves.

In fairness, Professor Piety seeks to limit the scope of her attack on the dignitary basis for free speech by pointing out, correctly I believe, that it borders on the absurd to treat commercial hawkers and corporate speakers obsessed with short-term profit as genuinely dignitary speakers.[65] Kant would roll over in his grave. Unfortunately, though, her response to the Court's argument that commercial and corporate speech is not about protecting dignitary speakers, but about preserving the dignitary and instrumental rights of hearers to receive uncensored information, is to infantilize hearers as incapable of coping with the commercial or corporate barrage.[66] It is, however, not necessary to assault the dignitary status of hearers to argue that, in the absence of a dignitary speaker, there is no dignitary value in being bombarded by false and misleading commercial information, or being subjected to one-sided presentations of important public and electoral issues merely because one side has an overwhelming economic advantage. In short, I believe that Professor Piety could have anchored her argument against extending commercial speech protection to false and misleading speech, and her argument for limiting the power of corporate America to dominate our political discourse, without demeaning the capacity of hearers to function as autonomous individuals. I wish that she had adopted a hearer-centered vision of commercial and corporate speech that views hearers with greater respect, and invokes that respect to place limits on the ability of non-protected speakers to trifle with the dignity of their hearers by lying to them about a commercial product.[67]

Second, Professor Piety argues that in institutions like elections, the market and scientific inquiry do not necessarily function better in settings where information flow is wholly uncontrolled.[68] She points out that false or misleading commercial speech can—and will—distort any economic market

64. "False consciousness" is a Marxist epithet for a mistaken belief by the masses about what is good for them. JOHN TORRANCE, KARL MARX'S THEORY OF IDEAS 5 (1995).

65. PIETY, *supra* note 35, at 79.

66. *Id.* at 132–34.

67. I attempt to describe the contours of such a hearer-centered First Amendment in Burt Neuborne, *The First Amendment and Government Regulation of Capital Markets*, 55 BROOK. L. REV. 5 (1989).

68. PIETY, *supra* note 35, at 165–66.

by leading rational, autonomous hearers to make inefficient judgments.[69] Frankly, given the wholly hearer-centered nature of commercial speech, that is all she needs to sustain her principal thesis that commercial speech protection should not be extended to false and misleading speech and speech about unlawful behavior.

Professor Piety also argues that a steady diet of one-sided profit-driven speech by corporations risks tilting elections unfairly by providing hearers with a misleadingly one-sided picture of complex issues.[70] Once again, lacking a dignitary speaker, Professor Piety could have based her critique on the affront to a hearer's dignitary interest in exercising fully-informed free choice that is caused by subjecting hearers to a sustained one-sided barrage of profit-driven corporate speech. That is what Justice Marshall did in *Austin v. Michigan State Chamber of Commerce*.[71]

Professor Piety argues, as well, that as an instrumental matter, invoking the First Amendment to permit a handful of giant media corporations to control the nation's book publishers, newspapers, and television outlets cannot be good for diversity of views or for innovative speech that pushes the envelope.[72] Once again, it's possible to argue that it is an affront to the dignity of hearers to subject them to such monolithic sources of information.[73]

Finally, Professor Piety points out that bad science, sometimes funded by interested profit-seeking corporations, can adversely impact everyone else's research, and will wreak havoc with a hearer's dignitary effort to construct rational preferences.[74]

As with her critique of autonomy, although Professor Piety confines her argument about the instrumental risks of uncontrolled speech to commercial and corporate speech, her critique inevitably bleeds into the political and

---

69. *Id.* at 189–90.

70. *Id.* at 166–67.

71. *See* 494 U.S. 652, 668 (1990) (identifying as a "serious danger the possibility that corporate political expenditures will undermine the integrity of the political process"), *overruled by* Citizens United v. FEC, 558 U.S. 310 (2010).

72. *See, e.g.*, PIETY, *supra* note 35, at 65 (writing that "[m]arket influences actually provide structural amplification, not for truth, but for ideas that are already popular, palatable, or attractive" and that "[b]ecause the access to means of communication is tied to financial means, commercial expression also inevitably results in amplification of the views congenial to the largest businesses"); *id.* at 68 (suggesting that "large institutions, which already have so many ways to control the news, end up getting their positions heard, while the public gets pushed to the side" (quoting Ben Casselman, *Three Stories a Day? How Young Reporters Learn to Skim*, COLUM. JOURNALISM REV., May/June 2004, at 65)).

73. *See generally* C. EDWIN BAKER, JR., MEDIA CONCENTRATION AND DEMOCRACY: WHY OWNERSHIP MATTERS (2007) (critiquing the concentration of mass media ownership and arguing that ownership dispersal would safeguard against abuses of media power and would more democratically distribute communicative power). The growth of the Internet and new forms of media complicates the argument that we are being harmed by information oligopolies.

74. Her description of Vern Countryman's confusion over the risk of cigarette smoking caused by false science is particularly chilling. PIETY, *supra* note 35, at 128.

artistic sphere. Taken to its logical conclusion, her picture of a systemically malfunctioning information market inhabited by gullible hearers and rapacious speakers argues for the restoration of the "bad tendency" test. Once again, I wish she had based her instrumental critique on the risk of harming rational, exogenous hearers.

Finally, Professor Piety seeks to rebut the cautionary argument against government censorship. She notes, briefly, that much of the fear-mongering about the risk of government speech regulation is premised on the actions of authoritarian regimes and institutions that lack democratic checks and balances.[75] Her principal response to the cautionary argument, though, is to note that democratic government is no worse (and is probably preferable) as a censor than profit-driven concentrations of private power like corporations and private employers.[76] She is right, of course, in observing that traffic in Mr. Madison's neighborhood must be managed by someone or something. She points out that under current deregulatory First Amendment rules much of the speech traffic management is concentrated in a small number of giant corporations and private employers. She argues that hearers would be better off if the speech traffic were managed by the government.

I fear that Professor Piety underestimates the power of the cautionary argument against government censorship. In assessing the relative danger of governmental as opposed to private censorship, it is of course true that private censors can cause real harm. Witness the McCarthy-era blacklist. It is also true that it is a form of "tough love" to subject hearers to much of what passes as commercial and corporate speech. But no private censor can put you in jail, or in a mental hospital, or take away your property, or exercise coercive controls over what you and your neighbors may read and hear. In my view, the potential for majoritarian suppression of weak voices, or partisan manipulation of information to stay in power, is simply too great to ignore, even in commercial and corporate settings. Removing some constraints on government regulation, perhaps by limiting the ability of powerful conduits to claim full-scale speaker protection, is worth thinking about. But taking a hammer to the cautionary argument against government regulation of speech, even commercial or corporate speech, seems, to me, to pose unacceptable risks.

In assessing the persuasive nature of Professor Piety's book, it is fair, I think, to divide her project into its three principal components: (1) an argument against extending constitutional protection to false or misleading commercial speech; (2) an effort to define commercial speech very broadly to permit government regulation of virtually all speech motivated by a short-term economic motive; and (3) an assault on the idea that corporations can

75. *Id.* at 224.
76. *Id.* at 135.

possess free speech rights at all, whether in the commercial, issue-oriented, or electoral spheres.[77]

The first goal—preventing the spread of First Amendment constitutional protection to false and misleading commercial speech—is something of a straw man. It is true that one or more Supreme Court Justices advocate such an expansion.[78] It's also true that a number of academics have questioned the two-tier First Amendment approach to commercial and noncommercial speech.[79] But, given the rationale of *Virginia Pharmacy* and *Central Hudson*, it would take an earthquake to move a majority of the Court to recognize that false and misleading commercial speech is entitled to full First Amendment protection.[80]

Given the lack of a dignitary speaker and the exclusively instrumental defense of the role of truthful commercial speech in improving the efficiency of the market and enhancing consumer choice, the Court would need an entirely new rationale for such an expansion. Such a rationale would have to depend on dignitary arguments at the level of the speaker, the hearer, or both. There is, however, no hint in either the commercial- or corporate-speech cases of an effort to imbue corporate speakers with human dignity. Corporate-speech rights are wholly derivative of the hearers' right to receive the information. It would, I believe, be awfully difficult to mount a convincing case that respect for the dignity of a consumer includes the right to be lied to about a product she is thinking of buying.

It is, of course, true that no person's First Amendment life or property is safe while this Court sits. The road to *Citizens United* demonstrates that

---

77. There is an occasional suggestion in the book that truthful commercial speech should not be protected at all, especially speech that subjects hearers to nonconsensual Madison Avenue barrages, but the issue is not fully developed. *Id.* at 137. The idea of a general "heckler's veto" in the commercial area seems much too broad. Targeted regulations that permit hearers to cut off unwanted commercial speech aimed at them are one thing. Broad-based regulation that cuts off truthful commercial speech to everyone is a much more difficult idea to justify. I'm not sure which version Professor Piety recommends.

78. Justice Thomas has rejected on several occasions the assumption that commercial speech is not entitled to the full protection of the First Amendment. *See* 44 Liquormart, Inc. v. Rhode Island, 517 U.S. 484, 518, 521–22 (1996) (Thomas, J., concurring in part and concurring in the judgment) (rejecting the "philosophical or historical basis for asserting that 'commercial' speech is of 'lower value' than 'noncommercial' speech"); Lorillard Tobacco Co. v. Reilly, 533 U.S. 525, 572 (2001) (Thomas, J., concurring in part and concurring in the judgment) (arguing that any government restrictions on speech should be subject to strict scrutiny, whether or not the speech is characterized as "commercial").

79. *See* Daniel A. Farber, *Commercial Speech and First Amendment Theory*, 74 NW. U. L. REV. 372, 386 (rejecting as unsound the judicial practice of distinguishing commercial speech based on economic motivation and subject matter for purposes of First Amendment analysis).

80. *Alvarez* is not to the contrary. First, there was a dignitary speaker in *Alvarez*. United States v. Alvarez, 132 S. Ct. 2537, 2542 (2012). Even chronic liars have human dignity. No dignitary speaker exists in a commercial speech setting. Moreover, although the Court declined to recognize that false speech is without any constitutional protection, a majority of the Court made it clear that false speech designed to induce a hearer to deliver a tangible benefit to the lying speaker is not protected. *Id.* at 2547.

precedent will not stop a runaway majority. But, the argument that the dignitary interests of hearers are respected by preserving their right to as much speech as possible on public issues, even from corporations, is a far cry from arguing that the dignity of a consumer is enhanced by demonstrably false or misleading speech about a product. Unlike the *Bellotti* or *Citizens United* contexts where truth about public policy is a subjective concept beyond the scope of government's ability to define, factually false statements about products are capable of objective assessment.

Not surprisingly, therefore, I believe that Professor Piety buries the argument for protecting false and misleading commercial speech, although I wish she had not trashed the ability of consumers to make rational choices along the way. It is precisely respect for the power of rational consumer choice in a free market that makes free speech protection for false and misleading commercial speech about a product such a nonstarter.

Professor Piety's second goal is to label as much economically motivated speech as possible as "commercial" in order to render it subject to government regulation for truthfulness. Corporate speech is her prime target. Her basic argument is that corporations speak and act only to further their short-term profit interests, stripping corporate speech of any dignitary value.[81] Without, or with minimal, speaker-based dignitary value, Professor Piety argues that corporate speech must pay an instrumental toll in order to claim first class privileged status.[82] False or misleading corporate speech, she argues, cannot pay such an instrumental toll, even when the subject is speech about public policy.[83]

To the extent she seeks to treat false, misleading, or unduly intrusive corporate speech about proposed economic transactions as commercial, Professor Piety is on strong and familiar ground. But when she argues that all corporate speech about public policy is commercial because it is economically motivated,[84] she crosses an indefensible line. For one thing,

---

81. *See* PIETY, *supra* note 35, at 148–50 (discussing how the corporate-profit motive affects speech).

82. *See, e.g.*, *id.* at 226–27 ("The question that must be answered before commercial speech is offered expansive protection is whether the purported benefits of such protection outweigh the harms of fewer restrictions. The evidence suggests they do not."); *id.* at 61 (arguing that various social costs associated with commercial expression outweigh the benefit of being able to use commercial speech "as material for self-expression"); *see also id.* at 57 (referring to the four general First Amendment interests identified by Emerson as "speaker-centered justifications").

83. *See, e.g.*, *id.* at 9 (suggesting that "'balance' [between the interests of the public and corporate interest in free expression] in the discussion of public concerns hardly seems to require insulation from liability for false statements, particularly false statements made in connection with commerce"); *id.* at 226 (arguing that the harms of fewer restrictions on corporate speech, including political and issue advertising, outweigh the benefits of greater protection).

84. *See, e.g.*, *id.* at 12 (advocating for expanding "the definition of 'commercial speech' to include everything that for-profit entities say, because no matter how it appears, no matter what communicative form it assumes, communications by for-profit entities are always and essentially promotional and hence 'commercial'"); *id.* at 31–32 (arguing that "the purpose of all commercial expression is promotion, including the sorts of press releases made by [a corporation] about its labor

the "economic motivation" test for defining commercial speech cuts much too broadly. Huge swatches of speech are economically motivated, including all speech by unions, speech by individuals with an economic stake in the outcome of a public policy debate, and the speech of all for-profit media. Moreover, unlike product advertising, when public policy speech is involved, government lacks the power to decide what is true or false. When public policy is at issue, hearers have a dignitary right to make up their own minds about the truth or falsity of speech about public issues, no matter who the speaker is. Vesting the government with power to pick and choose about the truth or falsity of speech about public policy is a cautionary nightmare. It empowers the local motorcycle gang to define truth.

Thus, to the extent Professor Piety seeks to expand commercial speech beyond its historic roots in product advertising to virtually all speech by corporations, she fails to persuade.

Professor Piety's third goal is her most ambitious. She attacks a corporation's First Amendment need to speak at all, even in noncommercial settings divorced from the electoral process.[85] Echoing Chief Justice Rehnquist's dissent in *Bellotti*,[86] she argues that, as a creature of the state, a corporation lacks First Amendment rights against the very government that has given it life. The Rehnquist dissent in *Bellotti* was a classic application of the principle that "the greater . . . includes the lesser."[87] Rehnquist argued that since the government is not obliged to create a corporation in the first place, it may place whatever restrictions it wishes on the final government-made product.[88] But the Rehnquist position runs into at least two roadblocks.

---

practices," and suggesting "that the format in which promotional speech is delivered does not affect its promotional character and therefore should be irrelevant for constitutional purposes"); *id.* at 35 ("Suffice it to say here that, as with issue advertising, the purpose of corporate, political advertising . . . is always, ultimately, to advance a commercial interest, because a corporation is an institution organized by law for a commercial purpose.").

85. *See, e.g.*, *id.* at 161 (suggesting that the corporate person "does not have a human need for self expression").

86. Justice Rehnquist made a similar argument in his opinion for the Court in *Posadas de Puerto Rico Associates v. Tourism Co. of Puerto Rico*, 478 U.S. 328, 346 (1986) (holding that the Puerto Rican government could restrict advertising for casino gambling because the government could have constitutionally banned gambling altogether).

87. *Id.* at 345–46.

88. For many years, such reasoning was a staple of the "right/privilege" dichotomy that empowered government to impose speech restrictions on the enjoyment of a "privilege," like public employment. *See, e.g.*, McAuliffe v. Mayor of New Bedford, 29 N.E. 517 (1892) (finding that the city of New Bedford was entitled to restrict a policeman's free speech rights as a reasonable employment condition). In recent years, the "right/privilege" dichotomy has been overtaken by the "unconstitutional conditions" doctrine limiting government's power to condition the enjoyment of a privilege on the waiver of First Amendment rights. *See, e.g.*, Speiser v. Randall, 357 U.S. 513, 528–29 (1958) (holding that veterans cannot be required to sign an oath curtailing their free speech as a condition for obtaining a tax exemption); Legal Servs. Co. v. Velazquez, 531 U.S. 533, 548–49 (2001) (holding that Congress cannot prohibit Legal Services Co. from funding organizations representing clients in an effort to amend or challenge existing welfare law, as this constitutes an unconstitutional restriction on private speech).

First, the train has long since left the station concerning the existence of corporate constitutional rights against the government. It is possible, of course, to rethink the idea that a corporation can be a "person" under the Fourteenth Amendment, but such a position would require overruling *Santa Clara Railroad* and 150 years of precedent.[89] Once corporations are recognized as "persons" under the Due Process Clause, it is impossible to argue persuasively that they are inherently incapable of enjoying First Amendment protection.

Second, once you move beyond false, misleading, or unduly intrusive commercial speech, while a corporate speaker may not qualify for dignitary status, hearers possess both significant dignitary and instrumental interests in hearing corporate speech about public issues. Thus, any effort at regulating such speech must demonstrate the usual compelling governmental interest. While I believe that such an interest exists in the electoral area in order to protect the egalitarian nature of the democratic process and prevent corruption, I do not perceive a serious argument in favor of censoring corporate speech about public issues across the board.

There is, however, a road not taken in Professor Piety's book that would justify regulation of corporate electoral speech.[90] Historically, the Supreme Court's recognition of a corporate constitutional right has always been a pragmatic judgment to vest corporate management with power to enforce rights enjoyed in common by the corporation's dispersed human constituents.[91] For example, everyone connected with a corporation engaged in operating a "press" shares a common interest in minimizing governmental interference with the product, and maximizing the ability to reach the largest possible audience. Not surprisingly, the Court has unhesitatingly recognized the First Amendment rights of the managers of press corporations to assert the First Amendment interests of the decentralized human members of the corporation.

Similarly, everyone connected with a corporation engaged in the sale of a product shares a common interest in providing potential consumers with truthful information about the product. Not surprisingly, the Court vests corporate management with the power to enforce the corporate community's shared interest in disseminating truthful commercial speech free from government censorship.

Finally, everyone connected with a nonprofit corporation formed to advance particular values shares an interest in maximizing the ability of the group to advance those values in an effective manner. Not surprisingly, the

89. *See* Santa Clara Cnty. v. S. Pac. R.R. Co., 118 U.S. 394, 396 (1886) (recognizing corporations as persons under the Fourteenth Amendment).

90. I discuss that road in Burt Neuborne, *Of "Singles" Without Baseball: Corporations as Frozen Relational Moments*, 64 RUTGERS L.J. 769, 774–76 (2012).

91. The classic article on the subject is John Dewey, *The Historic Background of Corporate Legal Personality*, 35 YALE L.J. 655, 658 (1926).

Court has recognized a First Amendment-based exception to government efforts to limit the speech of leaders of grassroots, ideological nonprofit corporations.

In settings, however, where conflicts of interest exist within the corporate universe over whether the behavior in question should take place at all, I believe that it would be a mistake to vest corporate management with the power to ignore the interests of significant components of the corporate universe. For example, the Court has refused to grant Fifth Amendment self-incrimination rights to a corporation because the exercise of such a right by corporate management would harm the interests of shareholders in learning about unlawful behavior by corporate agents.[92] The same reasoning should prevent recognition of a for-profit multishareholder corporate right to spend unlimited treasury funds to influence an election. Given the almost certain existence of political conflicts of interest within the corporate universe, it is, I believe, improper to vest corporate management with a constitutional right to use other peoples' money to advance management's personal political views.

Conclusion

Brandishing the First Amendment is a useful addition to the literature chronicling the negative consequences of a runaway First Amendment. Professor Piety makes a persuasive case for continuing to confine commercial speech protection to truthful commercial speech. She could, however, have built her case on the dignitary interest of rational and autonomous hearers. I fear that her two more adventurous goals—saddling virtually all corporate speech with limited commercial-speech protection; and questioning the right of corporations to speak at all—founder on the shoals of the dignitary right of hearers to receive uncensored information about the formation of public policy.

92. Hale v. Henkel, 201 U.S. 43, 69–70 (1906); *see also* Braswell v. United States, 487 U.S. 99, 119–20 (1988) (Kennedy, J., dissenting) (acknowledging that multishareholder corporations and other collective entities do not enjoy a Fifth Amendment right against self-incrimination, but arguing that single shareholder corporation should enjoy a right against self-incrimination).

Constitutional Adjudication, Free Expression, and the Fashionable Art of Corporation Bashing

BRANDISHING THE FIRST AMENDMENT: COMMERCIAL EXPRESSION IN AMERICA. By Tamara R. Piety. Ann Arbor, Michigan: University of Michigan Press, 2012. 342 pages. $70.00.

Reviewed by Martin H. Redish* & Peter B. Siegal**

I. Introduction

Late in 2011, Massachusetts Congressman James P. McGovern proposed a constitutional amendment to limit the terms "People, person, or citizens" as used in the Constitution to natural persons.[1] As to provisions that do not explicitly use the terms "People, person, or citizens," such as the First Amendment, the new amendment would clarify that "We the people who ordain and establish this Constitution intend the rights protected by this Constitution to be the rights of natural persons," with the goal and effect of rendering impossible any constitutional recognition of corporations.[2] Whatever one thinks about the merits of this proposal, there is little doubt that it taps into widespread confusion about and anger over the Supreme Court's holding in its 2010 decision in *Citizens United v. Federal Election Commission* that "the First Amendment does not allow political speech restrictions based on a speaker's corporate identity."[3] The widespread reaction of both legal scholars and educated lay people to the *Citizens United* decision was that it is preposterous to believe that a corporation could actually possess constitutional rights because a corporation is neither a "person" nor a "citizen."[4]

---

* Louis and Harriet Ancel Professor of Law and Public Policy, Northwestern University School of Law. The authors thank Vanessa Szalapski, Northwestern Law Class of 2014, for her valuable research assistance.

** A.B. 2008, University of Wisconsin; J.D. 2012, Northwestern University.

1. H.R.J. Res. 88, 112th Cong. § 2 (2011).

2. *Id.* §§ 1–2.

3. Citizens United v. FEC, 130 S. Ct. 876, 903 (2010). More limited ideas regarding possible constitutional amendments to overturn *Citizens United* have been advanced as well. *See* Lawrence Lessig, *Citizens Unite*, NEW REPUBLIC (Mar. 16, 2010), http://www.newrepublic.com/article/politics/citizens-unite# (proposing an amendment stating, "Nothing in this Constitution shall be construed to restrict the power to limit, though not to ban, campaign expenditures of non-citizens of the United States during the last 60 days before an election").

4. For an example of the near uniformly hostile reaction of free speech scholars to *Citizens United*, see MONEY, POLITICS, AND THE CONSTITUTION: BEYOND *CITIZENS UNITED* (Monica Youn ed., 2011). For a sampling of the hostility towards *Citizens United* reflected in the popular press, see, for example, Editorial, *The Court's Blow to Democracy*, N.Y. TIMES, Jan. 21, 2010, http://www.nytimes.com/2010/01/22/opinion/22fri1.html; Ed Crego et al., *Auctioning Off*

Most recently, the debate over corporate First Amendment rights has been impacted by the interesting and controversial—if seriously flawed—new book by Professor Tamara Piety, *Brandishing the First Amendment: Commercial Expression in America.*[5] Professor Piety's book develops an elaborate constitutional argument that all but excludes speech by profit-making corporations from the First Amendment's protective scope.

This widespread reaction, while perhaps politically understandable, reveals a complete lack of familiarity with well-established precepts of American constitutional law. In reality, the *Citizens United* Court's recognition of a corporation's ability to invoke constitutional rights was nothing new. Corporations have been invoking numerous constitutionalized and subconstitutionalized rights in court for many years.[6] Indeed, if Congressman McGovern's amendment ever managed to become law, one wonders how the provision's supporters would feel about the removal of the *New York Times* and *Washington Post*—both profit-making corporations, of course—from the First Amendment's protective reach.

Most of the battles over the constitutional status of corporations were long ago resolved in favor of allowing corporations to invoke constitutional guarantees. Today, corporate standing to challenge constitutional violations is so well established that it usually goes unnoticed. Corporations regularly invoke the Due Process Clause,[7] the Dormant Commerce Clause,[8] the Diversity Clause,[9] separation of powers protections,[10] and the Sixth and Seventh Amendment rights to jury trial.[11] Even when it comes to the First

---

Democracy, THE BLOG, HUFFINGTON POST (July 9, 2012, 10:43 AM), http://www.huffingtonpost.com/george-muno-frank-islam-and-ed-crego/citizens-united_b_1653556.html; Katrina vanden Heuvel, *A Court of, by and for the 1%*, WASH. POST, July 3, 2012, http://www.washingtonpost.com/opinions/roberts-court-is-still-a-conservative-defender-of-the-1-percent/2012/07/03/gJQA9xgLKW_story.html.

5. TAMARA PIETY, BRANDISHING THE FIRST AMENDMENT: COMMERCIAL EXPRESSION IN AMERICA (2012). Note that while Professor Piety's subtitle refers to commercial expression, she makes clear early on that she characterizes all expression by profit-making corporations as "commercial." *Id.* at 12–13.

6. *See* discussion *infra* Part II.

7. U.S. CONST. amend. V; U.S. CONST. amend. XIV, § 1.

8. *See* U.S. CONST. art. I, § 8, cl. 3 (inferred from granting to Congress the power "to regulate Commerce with foreign Nations, and among the several States, and with the Indian Tribes"); *see, e.g.*, Pike v. Bruce Church, Inc., 397 U.S. 137, 146 (1970) (holding that "[i]f the Commerce Clause forbids a State to require work to be done within its jurisdiction to promote local employment, then surely it cannot permit a State to require a person to go into a local packing business solely for the sake of enhancing the reputation of other producers within its borders").

9. U.S. CONST. art. III, § 2, cl. 1; *see, e.g.*, Hertz Corp. v. Friend, 130 S. Ct. 1181, 1192 (2010) (holding that for determining whether a federal court has diversity jurisdiction over a case with a corporate party, the court should look at the location of the corporation's "nerve center").

10. *See, e.g.*, N. Pipeline Constr. Co. v. Marathon Pipe Line Co., 458 U.S. 50 (1982) (determining, in a case between two companies, that the Constitution's separation of powers protection barred Congress from establishing legislative courts with complete jurisdiction over all matters arising under and related to bankruptcy law).

11. U.S. CONST. amend. VI, U.S. CONST. amend. VII; *see, e.g.*, S. Union Co. v. United States, 132 S. Ct. 2344, 2348–49 (2012) (holding that the Sixth Amendment provided a company with the

Amendment right of free expression, powerful corporate owners of newspapers and broadcast networks regularly invoke the First Amendment without the slightest controversy over their corporate form.[12] Moreover, since 1976, the Supreme Court has provided continually expanding First Amendment protection to commercial speech, which is invariably disseminated by profit-making corporations.[13]

Such practices should hardly come as a surprise. After all, if a corporation is defrauded in the marketplace by a contractor or competitor, would anyone seriously challenge that corporation's ability to resort to the judicial process to remedy the legal wrong done to it? Our economy would no doubt quickly degenerate into a state of chaos if corporations were denied the opportunity to vindicate their legal rights in court. But if no doubt exists that corporations have standing to vindicate *subconstitutional* rights and protections, how, purely as a logical matter, could they be categorically denied the opportunity to invoke the nation's highest law, the United States Constitution?

It is conceivable, we suppose, that one could acknowledge corporate rights to invoke *some* constitutional provisions, yet at the same time reject their ability to invoke the First Amendment right of free expression. It is certainly true that corporations have not been authorized to exercise *all* constitutional rights, especially in those situations in which it would be incoherent for them to do so. But it is far too late in the day to let the mere fact of their corporate form categorically disqualify them from constitutional protection. Those seeking to deny a particular constitutional right to corporate entities bear the burden of establishing such incoherence. Moreover, even within the confines of expressive rights, those who express shock and outrage at *Citizens United* would themselves readily extend those guarantees to the institutional corporate press without any sound basis for drawing so stark a distinction.[14]

Perhaps the problem is that the critics of *Citizens United* (and there are many of them) have failed to view the question of corporate free speech through the broader lens of constitutional theory. Once one grasps the reason why we have so readily extended so many other constitutional guarantees to corporations, it should be far easier to understand both why the values and

---

right to a jury trial for assessing significant criminal fines); Beacon Theatres, Inc. v. Westover, 359 U.S. 500, 501–12 (1959) (holding that a corporate plaintiff had the right to have a jury determine all issues of fact).

12. *See, e.g.*, Turner Broad. Sys., Inc. v. FCC, 512 U.S. 622, 636 (1994) ("There can be no disagreement on an initial premise: Cable programmers and cable operators engage in and transmit speech, and they are entitled to the protection of the speech and press provisions of the First Amendment."); United States v. Playboy Entm't Grp., 529 U.S. 803, 807 (2000) (discussing the Playboy Entertainment Group's First Amendment challenge to a statute that only applied to cable-television operators).

13. *See, e.g.*, 44 Liquormart, Inc. v. Rhode Island, 517 U.S. 484, 496–500 (1996) (summarizing important Supreme Court cases that expanded First Amendment protection for commercial speech).

14. *See infra* notes 60–62 and accompanying text.

purposes sought to be fostered by the First Amendment are significantly advanced by extending the provision's protections to corporate entities. This is true regardless of how one judges the moral, social, or economic impact of those entities.

In this essay we undertake three tasks. First, we explore the intersection between corporations and the Constitution, showing how widespread and well established that intersection is.[15] Second, we analyze critically the key arguments for categorically excluding corporations from the First Amendment's protective scope.[16] We do so primarily by intensively exploring and critiquing the most recent and important contribution to that side of the debate by Professor Piety.[17] Our critique finds serious flaws in each and every argument she advances.[18] But more importantly, our inquiry into Professor Piety's work enables us to glean from all the individual arguments a thematic failure to place the question of corporate First Amendment rights into the broader tapestry of constitutional theory, which has so readily and—with only very rare exception—consistently authorized corporations to invoke constitutional guarantees in court.

That brings us to the final task we take on here—namely, to fashion a coherent explanatory theory of constitutional adjudication in order to understand this widespread systemic choice in favor of extending the overwhelming number of constitutional rights and protections to corporations. To understand this decision about corporations and the Constitution, one must first understand why and how decisions are made to allow particular litigants to invoke constitutional protections. That inquiry leads us to a number of insights which, we believe, inexorably lead to the conclusion that corporations must be authorized to invoke the constitutional guarantee of free expression.[19] Thus, by grasping the fundamental premises of the theory of constitutional adjudication, we will be able to understand why corporations are authorized to invoke the First Amendment right of free expression.

We conclude that corporations do and should possess First Amendment rights. This is not necessarily because of the metaphysical nature of the corporate entity, the legal source of the corporation's existence, or the dictates of corporate theory. It is, rather, because of the vital instrumental role which the corporation serves in advancing the fundamental goals served by the First Amendment right of free expression through the process of private litigation.

In this important sense, authorizing corporate entities to invoke the right of free expression parallels numerous other instances in which corporations

15. *See* discussion *infra* Part II.
16. *See* discussion *infra* Part III.
17. *See* discussion *infra* Part III.
18. *See* discussion *infra* Part III.
19. *See* discussion *infra* Part IV.

have been authorized to invoke constitutional protections. Rather than focusing on the litigant-centric task of defining the corporation as either an aggregation of individuals bound together by contract or a goliath created by the state, courts should begin by identifying the values and policies that each individual constitutional provision seeks to advance. They should then ask whether authorizing injured corporations to invoke the provision's protections will foster and protect those policies and values. If the answer is in the affirmative, the corporation should be authorized to invoke the protection in judicial proceedings. This is so whether or not the corporation is itself the intended beneficiary of that provision. This approach wisely recognizes that incentivized litigants often perform an effective policing function, assuring that government complies with constitutionally imposed restraints and directives. In this sense, they act as an economically incentivized "private attorney general." And this is so even when the injured litigant is not itself the intended beneficiary of the provision it is enforcing. This approach to the theory of constitutional enforcement refuses to presuppose that protection of the litigant itself is the ultimate goal of the provision, because its focus is the systemic goal of checking government, rather than exclusively advancing the litigant's private interests. By describing and analyzing this instrumental view of the corporation's relationship to the Constitution, we hope to enable courts and scholars to understand what the Supreme Court has failed to explicitly state: that the important questions in corporate constitutional litigation are about what the Constitution has to say about preservation of our form of self-government, not about what corporate theory has to say about the Constitution.

While self-conscious abandonment of a narrow focus on corporate theory in constitutional cases leads to many of the results the Supreme Court has already reached, it can also correct errors that the Court has made. The Court has, for example, failed to recognize that corporations are "Citizens" for the purposes of the Privileges and Immunities Clause of Article IV,[20] despite the fact that that Clause was designed to limit the ability of individual states to begin trade wars by imposing economic hardship on out-of-staters.[21] As the primary engines of economic activity in the United States, corporations are particularly well-suited to enforce that provision, even though it may be troubling to consider them "citizens" in other contexts, such as with regard to the right to vote. The Court has apparently recognized this reality in the context of the Diversity Clause of Article III. That Clause sprung from the related belief that states might prejudice out-of-state entities, with negative consequences for interstate relations, and attempted to work around that threat by providing a neutral, federal forum in order to ensure that citizens of the Union could confidently transact business in foreign

20. U.S. CONST. art. IV, § 2, cl. 1.

21. *See* Mark P. Gergen, *The Selfish State and the Market*, 66 TEXAS L. REV. 1097, 1119 (1988) (discussing the history and original purposes of the Privileges and Immunities Clause).

states. If corporations can be "Citizens" for the purposes of the Diversity Clause, it is difficult to see why they cannot be "Citizens" under the Privileges and Immunities Clause. Although the Supreme Court has treated the two Clauses differently, the litigant-agnostic aim of interstate harmony should dramatically outweigh litigant-centric or anachronistically textualist stances on what sort of entity can qualify for recognition under those clauses.

From a broader perspective, the form of litigant-instrumentalism, which we advocate to rationalize corporate free speech rights, gains strength once one recognizes the fundamental role that constitutional adjudication plays in ensuring limited government. The goal of limiting the power and discretion of the political branches under the Constitution has led the Supreme Court to constrain government's treatment of individuals, even when enforcing those constraints requires granting windfalls to those who some may deem not particularly sympathetic guardians of constitutional rights. Thus, in litigation involving a number of provisions, the courts often sacrifice values as important as the accuracy of adjudicatory outcomes in order to protect the broader government-limiting aims of the Constitution. For example, the Fourth Amendment's exclusionary rule operates to protect only the victims of *successful* searches and seizures;[22] in doing so, it attempts to deter the use of unreasonable searches and seizures to obtain evidence despite the potentially significant impact that evidence might have on the resolution of the case. In excluding evidence obtained by illegal searches or seizures, the courts are effectively subordinating the aim of accuracy to the general aim of deterring constitutional violations.[23] Similarly, the broad protections of the *Miranda* rule show little regard for the sympathy due a particular defendant, who is often guilty of the crime of which he is accused; rather, the rule aims to regulate constitutionally pathological primary conduct.[24]

This government-limiting, regulatory-centric model of constitutional adjudication explains the extension of numerous constitutional protections to corporate entities. Nowhere is this more true than in the case of corporate free speech rights. Even if one doubts that corporations themselves are morally or conceptually deserving of that protection,[25] there is value in

22. *See* Mapp v. Ohio, 367 U.S. 643, 659 (1961) (discussing the ramifications of the exclusionary doctrine on criminal defendants).

23. *See, e.g.*, Richard A. Posner, *An Economic Approach to the Law of Evidence*, 51 STAN. L. REV. 1477, 1533 (1999) ("[Excluded] evidence is generally highly probative, and sometimes essential, and its exclusion has seemed a disproportionate sanction for police misconduct.").

24. *See id.* (explaining that "[t]he privilege [against self-incrimination] denies the court highly probative evidence, and the benefits of the privilege are exceedingly difficult to pin down" and the best argument in its favor is the "'strong policy in favor of government's leaving people alone'" (citing 8 JOHN HENRY WIGMORE, EVIDENCE IN TRIALS AT COMMON LAW § 2251, at 317 (John T. McNaughton ed., rev. ed. 1961))); *see also* Miranda v. Arizona, 384 U.S. 436 (1996) (creating what became known as *Miranda* rights).

25. For reasons to be explained, we do believe that a wholly litigant-centric model of constitutional adjudication does in fact justify corporate ability to invoke First Amendment rights. *See infra* subpart IV(B).

extending injured corporations the power to challenge constitutional violations. The marketplace of ideas, the ability of listeners to receive as much speech as possible, and the fear of governmental viewpoint discrimination or paternalism militate strongly against permitting the government to stifle corporate speakers. From a litigant-instrumental perspective, neither hostility to the corporation as a speaker nor a finding that the corporate speaker fails to benefit from expressive activities in the way that individual speakers do supports exclusion of corporations from the First Amendment's scope. Corporate expression may in numerous ways advance the constitutionally protected interests of others and further the regulatory-centric, government-limiting values underlying the expressive guarantee.

In the next section we discuss the numerous instances in which corporations have been authorized to invoke constitutional protections. In the section that follows we critically explore the wave of post-*Citizens United* scholarly arguments opposing First Amendment protection for corporate expression. We do so by focusing primarily on the theories developed by Professor Piety in her provocative, if seriously flawed, new book. We then suspend our disbelief about the arguments recognizing any speaker-based developmental or dignitary value in corporate speech. We demonstrate in this section that investing injured corporations with the power to challenge First Amendment violations nevertheless instrumentally serves important expressive values.

II. Corporations and Constitutional Adjudication

The First Amendment is one of many constitutional provisions whose enforcement is appropriately governed by an instrumental theory of constitutional adjudication. As a result, profit-making corporations have long been authorized to enforce those provisions through resort to the adjudicatory process. By briefly examining the adjudicatory dynamic underlying corporate enforcement of other constitutional provisions, we will be better able to understand the importance of the instrumental model as a rationale for recognition of corporate free speech rights.

A. The Corporation and Separation of Powers

The separation of powers provisions of the Constitution adopt what is essentially a prophylactic rule, or set of rules, protecting liberty.[26] The Framers structured the Constitution based on the recognition that, as Madison wrote in *The Federalist*, "[i]f angels were to govern men, neither external nor internal controls on government would be necessary."[27] But lacking the

26. *See* Martin H. Redish & Elizabeth J. Cisar, *"If Angels Were to Govern": The Need for Pragmatic Formalism in Separation of Powers Theory*, 41 DUKE L.J. 449, 451 (1991) ("[T]he Framers chose to rely on a number of different structural devices to check what they assumed to be the natural and inherent tendency of government to proceed toward tyranny.").

27. THE FEDERALIST NO. 51, at 319 (James Madison) (Clinton Rossiter ed., rev. ed. 2003).

requisite personnel for such an ideal form of government, the Framers enacted "formal, organized, and prophylactic structures to continually police potential abuses before the government is allowed to subvert liberty."[28] By separating government power, but creating enough overlap for each branch to exert a check on its coordinate branches, the Framers created a system in which the ambitions of each branch would serve to limit the excesses of the others.[29] Thus, the executive power is limited by the prospect of impeachment as well as Congress's control over the purse, the legislative power is limited by judicial review, and the judicial power is subject to impeachment and limited to the resolution of cases and controversies.[30]

The judiciary's ability to serve as a check on the coordinate branches of government is limited because, as an initial matter, the gears of the federal judicial machinery can be set in motion only in order to decide "cases or controversies."[31] And because the Supreme Court has unambiguously required a showing of "injury-in-fact" on the part of the plaintiff as a necessary element of the case-or-controversy requirement,[32] only one who has been injured by another's unlawful behavior can invoke the federal judicial process.[33] Thus, although the Framers undoubtedly considered judicial enforcement of the separation of powers to be a necessary protection against the tyrannical impulses of the other branches of government,[34] the judiciary itself cannot enforce the Constitution unless a potential litigant finds it to be in her best interest to sue.

---

28. MARTIN H. REDISH, MONEY TALKS: SPEECH, ECONOMIC POWER, AND THE VALUES OF DEMOCRACY 84 (2001).

29. Redish & Cisar, *supra* note 26, at 462–63; *see also* THE FEDERALIST NO. 51 (James Madison), *supra* note 27, at 319 ("Ambition must be made to counteract ambition.").

30. *See generally* JESSE H. CHOPER, JUDICIAL REVIEW AND THE NATIONAL POLITICAL PROCESS: A FUNCTIONAL RECONSIDERATION OF THE ROLE OF THE SUPREME COURT (1980) (examining the Supreme Court's role in relation to the other branches of government under the Constitution).

31. *See, e.g.*, Lea Brilmayer, *The Jurisprudence of Article III: Perspectives on the "Case or Controversy" Requirement*, 93 HARV. L. REV. 297, 297–98 (1979) ("[The case-or-controversy requirement] limits the jurisdiction of federal courts; when its requirements are not satisfied courts are without power to proceed.").

32. *See generally* Lujan v. Defenders of Wildlife, 504 U.S. 555 (1992) (holding that an environmental group lacked standing because they could not show any particularized injury-in-fact and thus the case was not an Article III case or controversy); Massachusetts v. Mellon, 262 U.S. 447 (1923) (holding that to meet the case-and-controversy requirement, a party challenging a federal statute must suffer direct injury from the statute rather than a general harm and that states may not challenge federal statutes on behalf of their citizens).

33. *See* Martin H. Redish, *The Adversary System, Democratic Theory, and the Constitutional Role of Self-Interest: The Tobacco Wars, 1953–1971*, 51 DEPAUL L. REV. 359, 381 n.62 (2001) (noting that the case-or-controversy requirement "embodies adversary theory" in that one may only advocate "one's self-interest" in federal court).

34. *See* THE FEDERALIST NO. 78 (Alexander Hamilton), *supra* note 27, at 466 (discussing the role of the judiciary and noting that "the courts were designed to be an intermediate body between the people and the legislature, in order, among other things, to keep the latter within the limits assigned to their authority").

It is well established that profit-making corporations qualify as such litigants. In the case of *Northern Pipeline Co. v. Marathon Pipe Line Co.*,[35] Marathon challenged the constitutionality of the Bankruptcy Act of 1978 on the grounds that it unconstitutionally conferred powers reserved to Article III judges on bankruptcy courts.[36] The case raised detailed questions regarding the breadth of the Article III power, the relationship between the courts and the political branches, and the scope of Congress's power under the Naturalization and Bankruptcy Clause.[37] In particular, the question was whether bankruptcy courts—staffed by non-Article III judges who lacked the protections of life tenure and fixed compensation—could be empowered to decide disputes arising under general state and federal law simply because such disputes might arise in or relate to cases under the bankruptcy laws.[38]

Relying expressly on the separation of powers—in particular, on the proposition that "[t]he Federal Judiciary was . . . designed by the Framers to stand independent of the Executive and Legislature—to maintain the checks and balances of the constitutional structure, and also to guarantee that the process of adjudication itself remained impartial"[39]—the Supreme Court held that the bankruptcy courts were powerless to adjudicate such disputes. In doing so, the Court did not even note the fact that the parties were corporations rather than natural citizens.[40] Although Marathon undoubtedly raised the unconstitutionality of the Bankruptcy Act out of its own self-interest—its victory led to the mandatory dismissal of a breach of contract action in which it was a defendant—its success served to protect the separation of powers and clarify the constitutional rules limiting Congress's ability to vest jurisdiction in non-Article III courts.

Because the *Northern Pipeline* Court completely ignored the corporate character of the parties, it cannot be said to have actually *held* that corporations may raise separation of powers arguments based on the requirements of Article III. But that is the point: The holding that an injured corporation has standing to object to a constitutional violation of the separation of powers was so well established that the Court took plaintiff's standing for granted. And no one appeared to disagree. Thus, the case represents a sterling example of the obvious and litigant-agnostic principle

35. 458 U.S. 50 (1982).

36. *Id.* at 56–57.

37. *See* U.S. CONST. art. I, § 8, cl. 4 (granting Congress the power "To establish an uniform Rule of Naturalization, and uniform Laws on the subject of Bankruptcies throughout the United States").

38. *See N. Pipeline*, 458 U.S. at 53–54 (discussing the challenged provisions of the Bankruptcy Act of 1978).

39. *Id.* at 58, 76 (plurality opinion).

40. Indeed, immediately upon introducing the parties, the majority noted that it would refer to petitioner Northern Pipeline Construction Co. as "Northern" and to respondent Marathon Pipe Line Co. as "Marathon." *Id.* at 56. Beyond those introductions, the opinion makes no reference to the corporate character of either of the parties.

that corporations, like individuals, can further constitutional aims by enforcing their personal self-interest.[41] That proposition applies to litigation in the context of a number of different constitutional provisions.

B. The Corporation and the Diversity Clause

The Diversity of Citizenship Clause of Article III extends the federal judicial power to suits between "citizens" of different states.[42] Though no corporation is a "citizen" in the purely literal sense of the term, it has long been understood that a suit between corporations from different states, or between a corporation from one state and an individual from another state, fall within the diversity jurisdiction of the federal courts.[43]

The reasons that corporations may appropriately be treated as citizens for purposes of the Diversity Clause are purely instrumental. Diversity jurisdiction was created to enable out-of-state citizens to avoid state courts, where they were likely to be subjected to prejudice in favor of in-state litigants. Such treatment would inevitably lead to retaliatory actions by courts of the other state, thereby creating substantial friction and disharmony between the two states. It was largely to avoid friction that the Clause was adopted. Corporations are today treated as "citizens" for purposes of the Clause, for the single reason that discrimination by state courts against out-of-state *corporations* amounts to discrimination against out-of-state *business*—the very type of interstate friction which the Framers feared. It is for these reasons that corporations may invoke the Clause even though textually it is confined to "citizens."

C. Seeking a Dividing Line

Some have sought to distinguish between the structural guarantees of the Constitution on the one hand and its rights-based provisions on the other hand, arguing that corporations should be permitted to assert claims and

---

41. It is worth noting the developing consensus that corporations are particularly effective litigants. For example, Professor Marc Galanter, who staunchly opposes the granting of constitutional recognition to corporations, observes that "[a]s law, driven by corporate expenditures, becomes more technical, complex, and expensive, individuals are just the wrong size to use legal services effectively." Marc Galanter, *Planet of the APs: Reflections on the Scale of Law and its Users*, 53 BUFF. L. REV. 1369, 1385–86 (2006). Similarly, Professor Gillian Hadfield, while lamenting the degree to which top-tier legal resources are devoted to litigation on behalf of corporations, notes that "the market for lawyers . . . overwhelmingly allocates legal resources to clients backed by corporate aggregations of wealth." Gillian K. Hadfield, *The Price of Law: How the Market for Lawyers Distorts the Justice System*, 98 MICH. L. REV. 953, 998 (2000) ("Driven by corporate demand, backed by corporate wealth, the legal system prices itself out of the reach of all individuals except those with a claim on corporate wealth."). As a practical matter, the societal interest in enlisting the most effective Hohfeldian litigants to do the bidding of broader society as private attorneys general is obvious.

42. U.S. CONST. art. III, § 2, cl.1.

43. *See* Louisville, Cincinnati, & Charleston R.R. v. Letson, 43 U.S. (2 How.) 497, 555 (1844) ("A corporation created by a state . . . seems to us to be a person, though an artificial one . . . and therefore entitled, for the purpose of suing and being sued, to be deemed a citizen of that state.").

defenses relating to governmental structure, but not to assert rights, which can belong only to individuals.[44] However, this line of argument ignores the strong instrumental justifications for allowing corporations to invoke constitutional protections which we previously explained. Even individually-held rights often have consequences, both theoretical and practical, well beyond the individual right-holder.[45] Thus, even if one were to reject the myth that corporations are reified beings divorced from their constituent parts, it would still not follow that the ability to make rights-based arguments should be withheld from corporations. To the contrary, granting corporations the protection of rights-based provisions can serve an instrumental purpose analogous to that served by the granting of structural protections to corporations.

Are there certain constitutional rights which corporations should not be authorized to invoke? The Supreme Court has consistently refused to recognize corporate rights to invoke a limited number of constitutional protections—for example, the Fifth Amendment right against self-incrimination[46] or the protection of Article IV's Privileges and Immunities Clause.[47] For reasons already explained, the Court's well established doctrine interpreting this provision of Article IV makes not the slightest bit of sense, and indeed, could well benefit from the lessons of instrumentalism we have delineated here. Perhaps the Fifth Amendment is the exception that proves the rule: one of those rare constitutional provisions where application to a corporate entity would be truly incoherent. But even if that were so, it would in no way undermine extension of constitutional protection to corporations in the numerous situations where the instrumentalist model fits.

III. The Attack on Corporate Free Speech Rights

In constitutional academic circles, corporation bashing has in recent years become a very fashionable activity. In particular, the Supreme Court's 2010 decision in *Citizens United*, which held that corporations have a First Amendment right to make independent expenditures for expressive purposes in political campaigns,[48] stimulated a wave of writings, both scholarly and popular, slamming the notion that corporations could possibly possess

---

44. *E.g.*, Burt Neuborne, *Felix Frankfurter's Revenge: An Accidental Democracy Built by Judges*, *in* MONEY, POLITICS, AND THE CONSTITUTION: BEYOND *CITIZENS UNITED*, *supra* note 4, at 195, 203–05.

45. *See* discussion *supra* subpart IV(A).

46. *See, e.g.*, United States v. White, 322 U.S. 694, 699 (1944) ("Since the privilege against self-incrimination is a purely personal one, it cannot be utilized by or on behalf of any organization, such as a corporation.").

47. *See, e.g.*, Blake v. McClung, 172 U.S. 239, 259 (1898) ("[A] corporation is not a citizen within the meaning of the constitutional provision that the citizens of each state shall be entitled to all the privileges and immunities of citizens in the several states." (internal quotation marks and citations omitted)).

48. Citizens United v. FEC, 130 S. Ct. 876, 913 (2010).

constitutional rights of *any* kind, much less a First Amendment right to contribute to the discourse or exercise in political campaigns.[49] To be sure, scholarly criticism of the idea of corporate free speech rights is not entirely new.[50] But it is as if *Citizens United* unleashed a flood of previously pent up intellectual contempt for the linkage of corporations and the Constitution.

A. The Lack of First Amendment Value of Corporate Expression

The traditional attack on corporate constitutional rights has focused on corporations' inability to exercise the First Amendment right of free expression. Long ago, Professor C. Edwin Baker contended that corporations are incapable of asserting First Amendment free speech rights basically because they have no soul.[51] He argued that the exclusive rationale for First Amendment protection is as an exercise of the speaker's free will; because a corporation is little more than an artificial, state-created, robotic profit-maximizer, it cannot be deemed to have a free will to exercise.[52] Similar sentiments have, more recently, been expressed by Professor Burt Neuborne.[53] He points to

> [t]he privileged status of the for-profit business corporation as an artificial, state-created legal entity blessed with unlimited life, limited-liability, highly favorable techniques of acquiring, accumulating, and retaining vast wealth through economic transactions having nothing to do with politics, and animated by one, and only one purpose—making money in the relatively short-term.[54]

While he concedes that "it makes sense to vest corporations with constitutional protection against improper economic regulation," he describes as an "unsupported . . . jump [the vesting of] corporations with non-economic constitutional protections that flow from our respect for human dignity."[55] "Robots," he notes, "have no souls. Neither do for-profit business corporations."[56] Most recently, the "corporations-have-no-soul" argument

---

49. *See supra* note 4.

50. *See generally* C. EDWIN BAKER, HUMAN LIBERTY AND FREEDOM OF SPEECH (1989) (arguing for the use of the liberty theory to interpret the First Amendment and concluding that corporate speech does not deserve the same protection as individual speech). *See also* Randall P. Bezanson, *Institutional Speech*, 80 IOWA L. REV. 735, 739 (1995) (advocating that institutional speech be protected by a separate and distinct analytical framework than individual speech).

51. *See* BAKER, *supra* note 50, at 218–19 (pointing to the motivations of political commercial speech in analogizing that class of speech to commercial speakers rather than individuals).

52. *Id.*

53. It should be noted that while Professor Neuborne echoes Baker's argument that corporations are incapable of self-realization, he has on occasion acknowledged that recognition of corporate free speech rights may, in limited instances, serve other First Amendment values. *See infra* note 106 and accompanying text.

54. Burt Neuborne, *Felix Frankfurter's Revenge: An Accidental Democracy Built by Judges*, 35 N.Y.U. REV. L. & SOC. CHANGE 602, 656 (2011).

55. *Id.* at 657.

56. *Id.*

has been raised by Professor Tamara Piety in her new book.[57] Professor Piety argues that "[c]orporations do not have a 'self' to be actualized or affirmed."[58] The corporation, she suggests, "is a collective with no corporeal existence."[59] It is therefore incapable of deriving value from expressive acts.

The first point to note in response is that the absence of a soul has never prevented corporations from invoking constitutional protections. Corporations have regularly taken advantage of constitutional provisions, under such provisions as the Diversity of Citizenship Clause, the Dormant Commerce Clause, the Due Process Clauses, and the Equal Protection Clause. With the possible exception of the Dormant Commerce Clause, none of these provisions, as Professor Neuborne appears to assume, is tied directly to matters of economic regulation. Professor Piety, in expressing outrage at what she considers the aberrational nature of First Amendment protection for profit-making corporations, fails even to acknowledge this broader—and far more complex—constitutional context.

More importantly, even First Amendment rights have long been extended, without great controversy, to profit-making corporations. Media corporations have, since their inception, been deemed to possess full First Amendment protections. While Professor Piety mysteriously ignores this fact, Professor Neuborne at least acknowledges the point, though his attempts to distinguish this form of expression from the speech of non-media corporations are unsuccessful. He focuses on the fact that it is the Freedom of the Press Clause, rather than the Free Speech Clause, that extends First Amendment protection to these profit-making entities.[60] Though Professor Neuborne may well be correct in this assertion,[61] it is unclear why anything should turn on this fact. Neuborne reasons that "[t]he business of operating a 'press' (in the Founders' time, a printing press) is the only economic activity explicitly protected by the Constitution."[62] So is Neuborne suggesting that someone who prints a newspaper for purposes other than profit—for example, a publisher driven solely by ideological considerations—is to be denied protection under the Press Clause? We cannot conceive of such a conclusion. Alternatively, would Neuborne suggest that an individual who

57. *See generally* PIETY, *supra* note 5, at 141–51 (defining the corporate identity as lacking the touchstone emotions of individualism).

58. *Id.* at 59.

59. *Id.*; *see also id.* at 77 ("Nonliving creatures do not 'self-actualize.'").

60. Neuborne, *supra* note 54, at 658.

61. A plausible argument may be made, however, that he is incorrect. One may reasonably challenge the notion that broadcast media are protected by the Press Clause, since their communications are not written and therefore not part of the "press." *See* Tom A. Collins, *The Press Clause Construed in Context: The Journalists' Right of Access to Places*, 52 MO. L. REV. 751, 759 (1987) (noting it is consistent with history to conclude that "the press clause was written to emphasize that written words were to be protected to the same extent as oral speech"). At the very least, this fact renders dubious recognition of any formalized distinction between the two branches of the First Amendment's expressive protection.

62. Neuborne, *supra* note 54, at 657–58.

makes her living by giving speeches, and who is therefore relegated exclusively to the protections of the Speech Clause, is to be denied First Amendment protection because she speaks, at least in part, for profit making purposes? Surely, Professor Neuborne would recognize that both of these conclusions would be nonsense. Whether or not the expression is the outgrowth of a profession is, for First Amendment purposes, besides the point. The Press Clause, like the Speech Clause, is facially agnostic as to the profit-making motivation of the speaker. In any event, even if true, Neuborne's argument is irrelevant. The issue is not whether the First Amendment protects profit-motivated *speakers*, but rather whether it protects profit-making *corporations*. It is anachronistic to believe that the Framers intended the Press Clause to protect the writings of modern profit-making corporations, since such entities did not even evolve until the Jacksonian period, long after the First Amendment was enacted.[63] Thus to conclude that the Press Clause was designed to protect a "profession" does not necessarily imply that it should also be construed to reach artificial corporate entities.

The very basis for Neuborne's and Piety's almost categorical[64] rejection of First Amendment protection of corporate speech belies any exception for protection for the corporate press. A media corporation is just as much "an artificial, state-created legal entity blessed with unlimited life, limited-liability, highly favorable techniques of acquiring, accumulating, and retaining vast wealth through economic transactions" as any non-media corporation. It is similarly "animated by one, and only one, purpose—making money."[65] Recall also that Professor Piety based her rejection of corporate speech protection on the fact that corporations lack "a 'self' to be actualized or affirmed."[66] Every one of these characteristics—the very characteristics which supposedly disqualify corporations from speech protection—apply equally to media corporations. Additionally, like all other mega-corporations, media corporations possess enormous economic power, enabling them to overwhelm any expressive marketplace they decide to enter. If these facts disqualify corporate expression from furthering the goals of the First Amendment's Speech Clause, it remains a mystery why that disqualification fails to apply also to the Press Clause. Finally, scholars who see a distinction between media corporations and other corporate entities fail

63. *See* BRAY HAMMOND, SOVEREIGNTY AND AN EMPTY PURSE: BANKS AND POLITICS IN THE CIVIL WAR 15–16 (1970) (describing the shift from a primarily agrarian economy towards an industrial and financial one); *see also* RONALD E. SEAVOY, THE ORIGINS OF THE AMERICAN BUSINESS CORPORATION, 1784–1855: BROADENING THE CONCEPT OF PUBLIC SERVICE DURING INDUSTRIALIZATION 256 (1982) (stating that corporation growth represented "the economic aspect of the poli[tic]al and social forces that democratized the United States during the age of Jackson"); *id.* (noting that incorporation laws "helped equalize the opportunities to get rich").

64. Neuborne would protect truthful commercial speech by corporations. *See infra* note 106 and accompanying text. And while she is not nearly as clear on the point, Piety might as well.

65. Neuborne, *supra* note 54, at 656; *see also supra* notes 53–56 and accompanying text.

66. PIETY, *supra* note 5, at 59.

to explain why non-media corporations that decide to publish a quarterly magazine or run a blog are not, either conceptually or practically, appropriately categorized as "the press" when they engage in such media-like activities.

Although Neuborne asserts—without the slightest support in constitutional doctrine—that the level of protection given to speech and press differ,[67] at no point has Supreme Court doctrine extended greater protection to the press than it has to speech. To the contrary, the Court has held that the protections are fungible.[68] Doctrinally, the two protections have always been treated identically, leaving no basis on which to distinguish between them in their protection of the expression of profit-making corporations. Thus, the fact that media corporations fall under the protective umbrella of the Press Clause, rather than the Speech Clause, provides absolutely no basis on which to justify disparate treatment for non-media corporate speakers who derive protection from the Speech Clause. It is a distinction without a difference.

Professor Neuborne's rejection of corporate speech protection takes on an almost surreal quality when one recalls his longstanding recognition of First Amendment protection for non-media corporate commercial speech.[69] In his more recent work, to his credit, he acknowledges the seeming inconsistency, but attempts—unsuccessfully—to rationalize it. Neuborne goes so far as to contend that the extension of First Amendment protection to commercial speech "not only fails to support a general right of corporate free speech, it cuts strongly against it."[70] This is because, in his words, "Commercial free speech is avowedly designed to maximize the economic efficiency of the market. As such, it is closely linked with the other constitutional protections afforded corporations in order to permit them to fulfill their economic mandate."[71] But rather than support his distinction between commercial and noncommercial corporate speech, Neuborne's argument actually proves the exact opposite. The level of constitutional protection extended to commercial speech far exceeds any protections extended to purely non-expressive economic activity under the Fifth or Fourteenth Amendment's economic substantive due process protections. Even Neuborne would have to acknowledge that over at least the last sixteen years the level of protection extended to commercial speech under the First Amendment far exceeds anything given non-expressive commercial activity

67. He asserts that the press receives "heightened" protection. Neuborne, *supra* note 54, at 658.

68. *See generally, e.g.*, Branzburg v. Hayes, 408 U.S. 665 (1972) (comparing the First Amendment protections afforded to a newspaper reporter to those afforded to ordinary citizens).

69. *See generally* Burt Neuborne, *The First Amendment and Government Regulation of Capital Markets*, 55 BROOK. L. REV. 5 (1989) (discussing the protections afforded to corporations and other commercial entities, particularly banks and other financial institutions subject to regulation by the Securities and Exchange Commission).

70. Neuborne, *supra* note 54, at 658.

71. *Id.* (footnotes omitted).

under the substantive due process heading.[72] This dramatic difference in the level of constitutional protection for commercial speech on the one hand and non-expressive commercial activity on the other hand undermines the notion that the rationale for the extension of constitutional protection is some generic concern with economic efficiency. To the contrary, something more than economic efficiency is at stake in the case of commercial speech, explaining the significant level of constitutional protection. The Court has concluded that commercial *speech* is deserving of substantial protection under the First Amendment, rather than under the minimal constitutional protection afforded commercial *conduct.*

It is true, as Neuborne asserts, that under current doctrine "commercial speech may be regulated in ways that would never be permitted in the first class speech compartment—most importantly on grounds of its falsity or misleading nature."[73] But that distinction flows not from the corporate nature of the speaker but rather from the assumption that the type of expression is of less value.[74] Indeed, while extending a lesser level of protection to commercial speech, the Supreme Court long ago emphasized that corporations nevertheless retain the "full panoply" of First Amendment rights to comment on political issues.[75]

In any event, commercial speech doctrine is a work in progress. In the relatively short period of thirty-five years since it was established, the level of First Amendment protection extended to commercial speech has gone from no protection to limited protection to substantial protection. Indeed, the government has not won a case challenging suppression of commercial speech in the Supreme Court in over twenty years, and in its most recent statement on the issue the Court at least implied that any regulatory distinction of expression premised on the commercial nature of the speaker deserves strict scrutiny.[76] In any event, the fact that the Court has not yet explicitly taken the final step of extending full First Amendment protection to commercial speech does not mean that the constitutional protection extended to it is fungible with the all but non-existent substantive due process protection extended to non-expressive commercial activity. Thus, the First Amendment's protection of commercial speech is—as a doctrinal matter, at least—clearly inconsistent with the sweeping Neuborne/Piety

---

72. *Compare generally, e.g.*, Thomson v. Western States Med. Ctr., 535 U.S. 357 (2002), *and* Lorillard Tobacco Co. v. Reilly, 533 U.S. 525 (2001) (both extending substantial First Amendment protection to commercial speech), *with, e.g.*, Ferguson v. Skrupa, 372 U.S. 726 (1963) (providing broad discretion to legislative bodies in regulating non-expressive commercial activity).

73. Neuborne, *supra* note 54, at 658.

74. *See, e.g.*, Ohralik v. Ohio State Bar Ass'n, 436 U.S. 447, 456 (1978) (affording "commercial speech a limited measure of protection, commensurate with its subordinate position in the scale of First Amendment values").

75. Bolger v. Youngs Drug Prods. Corp., 463 U.S. 60, 68 (1983) (noting that a "company has the full panoply of protections available to its direct comment on public issues").

76. Sorrell v. IMS Health, Inc., 131 S. Ct. 2653, 2664 (2011) (noting that content-based restrictions on speech require heightened scrutiny, and that "[c]ommercial speech is no exception").

argument that because corporations are soulless robots they should be denied substantial First Amendment protection.

Finally, it is important to note that scholars who dismiss constitutional protection for corporate speech on the grounds that corporations are incapable of self-realization are viewing the self-realization process in far too truncated a manner. They ignore the fact that the corporate form was developed for the very purpose of facilitating the self-realization of the individuals who formed corporations, who work for them, and who invest in them. In this sense, corporations are appropriately viewed as *catalysts* in the process of self-realization.[77]

B. The Inherently Harmful Nature of Corporate Speech

The arguments against First Amendment protection for profit-making corporations are not confined merely to the absence of expressive value of corporate speech. Professor Piety in particular devotes much of her recently published book to arguments focused on the inherent harm caused by corporate speech. For one thing, she asserts, corporate speech may often manipulate consumers into making unwise or unnecessary choices in the marketplace, because consumers have been shown to be psychologically vulnerable to such manipulation.[78] In addition, she contends that corporate speech will often lead to economic instability.[79] Furthermore, she asserts, corporate expression may give rise to severe "social costs," such as "in environmental pollution or child labor."[80] In a world of free and open corporate speech, she suggests that "social reality is fairly relentlessly focused on the material. The billions of dollars spent on marketing face little competition from speakers on other issues: political, religious, spiritual, altruistic, or, most pointedly, antimaterialist."[81] She suggests that "[c]omparatively little time is given in the commercial world to issues of labor, family, religion, altruism, or other noncommericial aspects of life, except insofar as they can be translated into something for sale."[82] Professor Piety here makes a stirring argument against the value choices implicitly or explicitly advocated by much corporate speech. But it is surprising that she believes these arguments have any role to play in the debate over the level of First Amendment protection to be extended to corporate speech. To the contrary, her arguments represent a paradigmatic illustration of viewpoint

77. For detailed development of this argument, see REDISH, *supra* note 28, at 76–80.

78. PIETY, *supra* note 5, at 107–20.

79. *Id.* at 186–87.

80. *Id.* at 61.

81. *Id.* at 60–61; *see also id.* at 228 ("Commercial expression also plays a role in boosting demand for products produced in a manner some find objectionable, whether because of animal testing, labor practices, or support for groups or causes offensive to some or many people. . . . [I]t also reinforces gender and racial stereotypes that undermine other societal attempts to remove barriers to full equality for women and disadvantaged minority groups.").

82. *Id.* at 61.

discrimination—a mode of analysis universally shunned in First Amendment doctrine and theory.[83] As one of us has previously argued,

> Viewpoint discrimination . . . flows from normative premises determined by . . . factors of political, social, economic, moral or religious beliefs or concerns that are wholly external to the First Amendment itself. They grow not out of the process-based analysis that seeks to create the most viable or appropriate constitutional system, but rather from unrelated personal beliefs of those imposing the restriction. To those seeking to impose viewpoint discrimination, the First Amendment is not something to be deciphered and structured, but rather a potential obstacle to attainment of their political or ideological values and goals that needs to be circumvented.[84]

Professor Piety effectively employs corporate speech as a surrogate for all of the sociopolitical views which she detests. Professor Piety's argument for upholding the constitutionality of the suppression of corporate speech is that it uniformly advances a set of values which she deems offensive. While we appreciate her candor and admire the fervency of her moral beliefs, we must categorically reject the viewpoint-based nature of her constitutional argument.

Professor Piety, in short, displays a mystifying willingness to fall back on naked viewpoint offensiveness as one of her primary rationales for denying corporate speech First Amendment protection. She also inexplicably (and inexcusably) fails even to discuss, much less distinguish, the fact of First Amendment protection for the speech of media corporations. These failures underscore the most significant failure in her analysis: her general failure to place the issue of First Amendment protection for corporate speech into a broader constitutional framework. The simple fact is that corporations have long been authorized to assert violations of numerous constitutional provisions and to invoke numerous constitutional protections.[85] This is not necessarily because we have made the ex ante determination that the corporations themselves will benefit from or flourish because of such protections. As is often the case in constitutional theory, corporations are in many instances permitted to assert these constitutional protections as litigants for the simple reason that we know that they will be motivated by economic considerations to protect and defend broader societal goals and values which are intended to be implemented by adoption of particular constitutional

---

83. *See, e.g.*, Schacht v. United States, 398 U.S. 58, 62–63 (1970) (holding unconstitutional a congressional ban on the unauthorized wearing of American military uniforms in a manner calculated to discredit the armed forces); Good News Club v. Milford Cent. Sch., 533 U.S. 98, 107 (2001) (holding that a school's exclusion of a Christian children's club from meeting after hours at school, based on its religious nature, is an unconstitutional viewpoint discrimination); Legal Servs. Corp. v. Velasquez, 531 U.S. 533, 549 (2001) (finding unconstitutional a restriction prohibiting funding to organizations representing clients seeking to challenge existing welfare laws).

84. Martin H. Redish, *Commercial Speech, First Amendment Intuitionism and the Twilight Zone of Viewpoint Discrimination*, 41 LOY. L.A. L. REV. 67, 113–14 (2007).

85. *See* discussion *supra* Part II.

provisions. The corporation itself, in this theory of constitutional adjudication, is viewed simply as a means to a broader end. Because Piety fails even to attempt to place the issue of corporate speech protection within the broader theoretical framework underlying constitutional adjudication, she fails to grasp the real reasons why corporations must be permitted to invoke First Amendment protections. It is to this question that our analysis now turns.

IV. Instrumental Justifications for Constitutional Protection of Corporate Expression

Constitutional concerns having nothing to do with the corporation per se militate strongly in favor of allowing corporations to raise constitutional arguments. The Reporters are filled with cases in which litigants have advanced self-interested constitutional arguments, the success of which simultaneously furthered the aims of those litigants as well as aims that extend far beyond the interests of the litigants themselves. It is in this sense that the corporation can be viewed as a type of self-interested "private Attorney General," advancing its own interests with the simultaneous collateral benefit of furthering values enshrined in the Constitution.[86] And when considered in light of the ability of the corporation to use its status as a litigant to enforce broader societal norms, the question of whether a corporation should be entitled to invoke a given provision requires reference not to any dogmatic view of corporate theory, but rather to the systemic concerns of the constitutional provision sought to be enforced. It is in this sense that constitutional litigation is a largely instrumental process.

A. The Corporation as a Hohfeldian Plaintiff

Although the American legal system is undoubtedly concerned with achieving desirable outcomes for society as a whole,[87] the only mechanism by which courts may permissibly engage in the pursuit of the public good is by deciding discrete cases which resolve claims brought by self-interested litigants.[88] This is the essence of the nation's "private rights" model of adjudication.[89] In the words of a leading commentator, "[I]t is still holy writ that the citizen *qua* citizen is not a proper party plaintiff in a lawsuit testing

86. *See* Associated Indus. of N.Y. State, Inc. v. Ickes, 134 F.2d 694, 704 (2d Cir. 1943) (coining the term "private Attorneys General"); *see also* Martin H. Redish, *Class Actions and the Democratic Difficulty: Rethinking the Intersection of Private Litigation and Public Goals*, 2003 U. CHI. LEGAL F. 71, 87 n.60 (discussing the "private Attorney General" concept).

87. *See* Redish, *supra* note 86, at 86 (rejecting the notion "that federal adjudication is incapable of advancing social, economic, or political interests that extend well beyond the personal interest of the individual litigant").

88. *See supra* notes 31–33.

89. For a description and critical analysis of the private rights adjudicatory model, see MARTIN H. REDISH, THE FEDERAL COURTS IN THE POLITICAL ORDER: JUDICIAL JURISDICTION AND AMERICAN POLITICAL THEORY 87–109 (1991).

questions of constitutionality."[90] Rather, in order to rule on the constitutionality of a governmental act, a court must first find that the plaintiff "is seeking a determination that he has a right, a privilege, an immunity, or a power."[91] Such litigants are often referred to as "Hohfeldian" plaintiffs, because the concept of the self-interested litigant was first articulated by Professor Wesley Hohfeld.

That the adjudicatory process delegates to private litigants the responsibility for fostering and protecting constitutionally protected interests should come as no surprise. Both the Constitution and the statute books abound with private rights designed to enlist personally aggrieved litigants in the broader effort to enforce systemic legislative and constitutional aims. As one of us has previously asserted, "[p]rivate litigation may often do the government's work for it, by deterring and punishing violations of law."[92] But while many have criticized certain subconstitutional regimes which rely on private enforcement to ferret out wrongdoing,[93] there can be little doubt that litigants who seek to vindicate their own self-interest by advancing constitutional arguments enable the courts to further established public policies without raising the difficulties that attend other private attorney general actions.

It is true that private litigants who bring compensatory suits intend primarily to vindicate private interests.[94] Although private compensatory litigants may of course seek to advance broader societal goals as an incident to vindication of their private rights,[95] their primary (and often exclusive) personal goal is to use the legal system to advance their own private rights. But in pursuing their own private interests, they will often incidentally protect and enforce the systemic interests that underlie the substantive law they seek to enforce.

The profit-making corporation, by its very nature, falls into this category: by the rules of corporate law, the corporation must function as a

---

90. Louis L. Jaffe, *The Citizen as Litigant in Public Actions: The Non-Hohfeldian or Ideological Plaintiff*, 116 U. PA. L. REV. 1033, 1033 (1968).

91. *Id.* For the origins of the phrase "Hohfeldian" turn to Wesley Newcomb Hohfeld, *Some Fundamental Legal Conceptions as Applied in Judicial Reasoning*, 23 YALE L.J. 16 (1913).

92. Redish, *supra* note 86, at 86.

93. A number of scholars, for example, have argued that large class actions—the prototypical examples of "private attorney general" suits—create conflicts of interest between attorneys and class members. *See, e.g.*, John C. Coffee, Jr., *Understanding the Plaintiff's Attorney: the Implications of Economic Theory for Private Enforcement of Law Through Class and Derivative Actions*, 86 COLUM. L. REV. 669, 671–72 (1986) (discussing misincentives that attorneys face during a class action suit). One of us has argued in Redish, *supra* note 86, at 73, that Federal Rule of Civil Procedure 23, which authorizes class actions, undermines democratic legitimacy by using a procedural device to drastically and fundamentally alter substantive law.

94. Redish, *supra* note 86, at 85–86.

95. *Id.*

self-interested litigant.[96] It is, for that reason, perhaps the ideal private attorney general: the benefits of corporate litigation necessarily accrue to shareholders via changes in stock price, and when a corporation sues, there is *always* a real class of plaintiffs—the shareholders—whose decision to opt into the corporate structure necessarily confers on corporate managers a right to sue on their behalf.[97] The corporation engaged in constitutional litigation, in other words, can realize many of the benefits of the class action without falling victim to its weaknesses.[98]

B. Corporations as First Amendment Private Attorneys General

In deciding whether to allow corporations to invoke constitutionalized protections in court, it makes little sense—as opponents of corporate speech protections have—to obsess over whether a corporation possesses a soul, is capable of self-realization, or is conceptually characterizable as a "person."[99] The question, to be asked, rather, is whether enabling corporations to invoke First Amendment protection advances the First Amendment ball from Point A to Point B. In other words, the inquiry is largely an instrumental one: Will the corporation be in a position to advance or protect the values sought to be fostered by our constitutional commitment to the principle of free expression?

Of course, if one believes that the *only* value served by the constitutional protection of free expression is the self-development of the speaker, one might conclude that corporate entities are undeserving of that protection. Even on the basis of that narrow assumption, however, for reasons previously discussed, one should recognize that the corporation performs an important catalytic function in fostering the self-development of the individuals who create and participate in the operation of the corporation.[100] But the main problem with this view is that it posits an unduly narrow version of the role properly served by litigation brought to vindicate First Amendment rights. Whether or not corporations self-realize as speakers, by advocating on behalf of the First Amendment they

---

96. To the extent that the corporation must act to maximize shareholder wealth, directors are required to initiate suits and advance legal arguments only to the extent that such suits and arguments would benefit shareholders. While it is of course plausible that agency problems might lead managers to authorize suits and legal arguments with the aim of benefiting the public interest as distinct from the interests of shareholders—in other words, to engage in lone ranger litigation—such a course of action would reflect a violation of managerial responsibility to shareholders.

97. *See, e.g.*, HERBERT HOVENKAMP, ENTERPRISE AND AMERICAN LAW, 1836–1937, at 15–16 (1991) (discussing the historical role of the corporation in initiating legal action on its own—and thus its shareholders'—behalf).

98. *See* Redish, *supra* note 86, at 77 (arguing that "faux" class actions—those in which suit is not, "in any realistic sense, brought either by or on behalf of the class members"—improperly transform statutory provisions from guarantors of private rights to grants of a roving right to ferret out wrongdoing).

99. *See* discussion *supra* subpart III(A).

100. *See* discussion *supra* Part I.

instrumentally advance and protect broader systemic interests which underlie the constitutional commitment to free expression.

This analysis supports the universal acceptance of the right of profit-seeking media corporations to invoke First Amendment protections. We do not extend them this protection in order to enable these corporations to self-realize, or because they are appropriately conceptualized as humans any more than any other corporation is. Nor do we extend them protection because media corporations are exercising the right of the *press*, rather than the right of *speech*—a distinction without a difference, as well established First Amendment doctrine makes clear.[101] We do so, rather, for two commonsense reasons: (1) to enable them to check governmental excesses through their publications or broadcasts, and (2) to assist readers and listeners in their own self-realization by providing them with information which may facilitate their personal and collective decision making processes.[102] This is an extremely important insight for purposes of the debate over the free-speech rights of non-media corporations. The simple fact is that non-media corporations are capable of performing the exact same instrumental functions on behalf of the values sought to be fostered by the constitutional commitment to the principle of free expression. Protecting the expression of non-media corporations—even assuming that all such expression is self-advancing but not self-developing—serves the dual functions of checking government and informing the public.

The point may be illustrated by use of a hypothetical example. Imagine that the federal government has enacted a law which provides that profit-making non-media corporations are permitted to spend treasury funds for expression that approves of or supports the president's policies, but that these corporations are not permitted to spend treasury funds for expression that criticizes or dissents from those policies. Presumably, those who believe that profit-making non-media corporations are not protected by the First Amendment would logically have to conclude that such a law is constitutional. After all, the law would in no way inhibit the speech of an actor protected by the First Amendment. But it is difficult to imagine any court upholding such a law against a First Amendment attack. On its face, the law constitutes a blatant form of selective viewpoint suppression, designed to skew public debate in favor of those in power. Yet all the arguments for excluding profit-making corporations would still apply: the speakers remain soulless, mindless, robotic profit maximizers, as well as artificially created centers of overwhelming economic power.[103] If corporations are inherently incapable of exercising First Amendment rights,

---

101. *See* discussion *supra* subpart III(A).

102. *See generally* Vince Blasi, *The Checking Value in First Amendment Theory*, AM. B. FOUND. RESEARCH J. 521 (1977) (describing one value of free expression as checking the abuse of official power).

103. *See* discussion *supra* subpart III(A).

it logically cannot matter what form the regulatory interference with corporate expression takes. In our hypothetical, however, the corporate nature of the victim should make little difference: we simply cannot allow government to skew public debate in this manner, thereby threatening core democratic values. By allowing corporations to challenge the constitutionality of this law, we would be policing governmental excess, thereby implementing the regulatory-centric model of constitutional adjudication.

By a process of reverse engineering, then, we should be able to glean from this example precepts of First Amendment theory that recognize what an instrumentally important role profit-making corporations can play in protecting and advancing important First Amendment values. Our hypothetical demonstrates that viewing the First Amendment from a purely speaker-centric perspective unduly truncates the safeguards necessary to assure that government acts in accordance with the social contract between government and citizen dictated by our governmental form. In addition, we need to add both listener-centric and regulatory-centric perspectives of free speech protection: there are certain behavioral patterns in which government may not be permitted to engage if we are to keep government within the confines of the First Amendment. Our hypothetical example is a perfect illustration of the importance of this regulatory-centric model of free speech protection. The danger may be assumed not to be to the self-development or self-realization of the corporation as speaker, if one has concluded that corporations are undeserving of such protection.[104] Nevertheless, it is vitally important to vest the corporation with standing to challenge this blatant violation of free expression, lest government be permitted to exert dangerously excessive power to skew the nature of public debate.

As previously noted, Professor Neuborne himself, in earlier writing, essentially grasped the concepts of the listener-centric and regulatory-centric models of First Amendment theory.[105] He once defended commercial speech protection, despite what he deemed to be the absence of speaker self-realization, both because of the First Amendment right of the listener to be informed and the fear of governmental paternalism inherently reflected in much governmental suppression of truthful commercial speech.[106] These concerns are strikingly similar to the listener-centric and regulatory-centric models which we have described.[107] Thus, Professor Neuborne has himself acknowledged that far more is at stake in First Amendment enforcement than

104. *But see* discussion *supra* Part IV(B) (discussing catalytic self-realization rationale for protecting corporate speech).

105. *See supra* notes 69–75 and accompanying text.

106. Neuborne, *supra* note 69, at 27. It should be noted that Professor Piety, unlike Professor Neuborne, appears to actually embrace the notion of governmental paternalism as a basis for rejecting or at least limiting corporate free speech rights. Piety, *supra* note 5, at 124–25.

107. *See* discussion *supra* Part I.

merely the developmental benefit to the speaker. Indeed, respected free speech philosophers have suggested that the *only* value served by the guarantee of free expression is listener-based.[108] While we believe that this perspective unduly truncates the category of those who benefit from free speech protection, surely more must be involved than a narrow focus on benefit to the speaker. Yet those who have attacked—indeed, mocked—*Citizens United* for its extension of First Amendment protection to mindless, soulless profit-making corporate automatons have either (as in the case of Professor Piety) completely ignored the nonspeaker benefits of free expression or (as in the case of Professor Neuborne) conveniently forgotten that on other occasions they themselves have recognized those very nonspeaker values which may be defended and protected as vigorously by a corporate speaker as by an individual one.

If scholars such as Professor Neuborne recognize that corporate speech fosters both the listener- and regulatory-centric models of free expression in commercial speech contexts, it is difficult to understand why those very same values are not fostered at least as effectively in the context of a political campaign by protecting corporate political speech. It is in just this context that many free speech scholars have long argued that the need for an informed public is at its greatest.[109] Shutting down the ability of corporations to contribute information to political campaigns undermines this listener-centric model as much as would the restriction of the expression of human speakers. It should not be forgotten, after all, that the expression sought to be suppressed in *Citizens United* was a movie critical of a major presidential candidate in the midst of a political campaign—expression which could potentially influence voter decision making. It is also in the political context that the need for the regulatory-centric model is arguably at its height, because the temptation to selectively suppress expression in support of the opposition is at its height in this context. In this context, it is worthy of note that the overwhelming majority of corporate sponsored expression is likely to support pro-capitalist policies and conservative candidates. Indeed, it is this very fact which has generally led to so much concern on the part of liberals about *Citizens United*. Yet it is this knowledge of the likely content of the suppressed expression that renders the push against corporate speech a thinly veiled form of wholly impermissible viewpoint regulation.[110]

---

108. *See, e.g.*, ALEXANDER MEIKLEJOHN, POLITICAL FREEDOM: THE CONSTITUTIONAL POWERS OF THE PEOPLE 27 (1965); Robert H. Bork, *Neutral Principles and Some First Amendment Problems*, 47 IND. L.J. 1, 24–26 (1971) (asserting that only one of four benefits from Brandeis's *Whitney* concurrence—"the discovery and spread of political truth"—can be held above other claimed freedoms (citing Whitney v. California, 274 U.S. 357, 375 (1927))).

109. *See* Bork, *supra* note 108, at 24–28 (arguing that only explicitly political speech should receive constitutional protection).

110. For a discussion of Professor Piety's overt reliance on viewpoint discriminatory factors in her case against corporate protection speech, see *supra* note 83 and accompanying text.

It does not necessarily follow, we should note, that *Citizens United* was correctly decided. One could arguably accept the premise of corporate free speech rights in the abstract, yet nevertheless conclude that the importance of limiting the role of money in political campaigns constitutes a sufficiently compelling interest to justify such restrictions. While this would not likely be our position, we do not reach the issue because the question of the generic role of money in political campaigns is beyond the scope of our present inquiry.[111] The point to be emphasized here is that if one were to reach this conclusion, it could not be grounded in the notion that corporations are incapable of exercising First Amendment rights, and therefore would logically have to apply equally to large campaign expenditures by individuals.

Of course, if one embraces, rather than rejects, the notion of governmental paternalism as grounds for regulating speech, then one would refuse to deem viewpoint-selective governmental behavior constitutionally troublesome. Professor Piety appears to do just that. Recall that she asserts as one basis for excluding corporate speech from the First Amendment's scope the danger that such expression will unduly manipulate the minds and behavior of consumers.[112] But her reasoning necessarily rejects the foundational democratic assumption that the people are able to judge the wisdom of competing arguments for themselves without "assistance" by government.

Piety argues that although

> [m]uch Western thought is deeply invested in this idealized notion of human cognition in which actions taken by impulse or under the influence of emotion are somehow corrupt or to be rejected in favor of deliberative actions. . . . This understanding of human cognition does not appear to be supported by the most recent research. This research suggests that human beings' capacity for rational behavior is subject to significant limitations, that we have bounded rationality.[113]

Therefore it is appropriate, she concludes, for government to suppress corporate speech that would bring about unwise consumer choices. If one were to follow Piety's logic to its ultimate conclusion, it would necessarily imply that *some* corporate speech would, in fact, have to be protected—any corporate speech which would lead to "wise" consumer conclusions. Presumably, some agency of government would be placed in the position of deciding which corporate speech would have to be protected and which would not by deciding which corporate speech would induce "wise" consumer choices. But if accepted, Piety's argument would effectively do away with the foundational premises underlying our democratic system, not

111. Also beyond the present inquiry is the question of constitutional protection for the anonymity of such expression.

112. *See supra* note 78 and accompanying text.

113. PIETY, *supra* note 5, at 83 (footnote omitted).

to mention the societal commitment to free expression. If the people are incapable of being trusted to make rational choices on the basis of free and open debate and therefore must be aided by selective—and paternalistic—governmental suppression, the inevitable conclusion must be that the entire democratic process cannot be trusted.

In any event, Piety's argument proves far too much. The kind of speech involved in *Citizens United* and many other cases has nothing to do with consumer choices in the commercial marketplace. It concerns, rather, debate over the political choices open to voters. If one applies her reasoning to this context (and there would appear no logical basis on which not to do so), then it is impossible to understand why corporate political speech is any more likely to bring about irrational choices than political appeals made by individuals. If the electorate is not to be trusted to make choices on the basis of free and open debate, it logically matters not at all who the speaker is. Thus, consistent with the theme of much of our criticism of her new book,[114] Professor Piety's failure even to attempt to place her stinging attack on corporate speech within either the broader context of democratic theory, the law and theory of free expression, or the doctrines of constitutional law which have long extended constitutional protections to corporations, ultimately deprives her theory of any persuasive force.

V. Conclusion

There appears to exist a post-*Citizens United* reflex among the uninformed and the ideologically driven to assume that because corporations are not humans, they are—both legally and metaphysically—incapable of asserting *any* constitutional right, much less the First Amendment right of free expression.[115] As we have made clear, however, such a view could not be further from the truth. Corporations have long been authorized to assert numerous constitutional provisions, including the First Amendment right of free expression. And there are very good reasons for such a practice. Regardless of how one views the corporation on a metaphysical level, enabling corporations to invoke the First Amendment right of free expression serves important instrumental purposes in fostering constitutional values and checking government.

Like many scholars before her, Professor Piety expresses intense political hostility towards the modern profit-making corporation, but ignores all of the doctrine and history surrounding corporate assertion of constitutional rights. She even ignores the simple fact that profit-making media corporations have long asserted First Amendment rights, and therefore fails even to attempt to distinguish the speech of non-media corporations. Instead, she relies primarily on arguments wholly inconsistent with well-

114. *See supra* subpart III(A).

115. *See supra* note 4.

established precepts of free speech theory—for example, political or social hostility to the ends sought to be achieved by corporate speech and skepticism about the public's ability to judge competing arguments.[116] Such scholarly work not only fails to advance free speech thought; if left unanswered it would set First Amendment theory back significantly.

116. *See supra* note 78–86 and accompanying text.

* * *

Essay

Changing the Litigation Game: An *Ex Ante* Perspective on Contractualized Procedures

Daphna Kapeliuk* & Alon Klement**

I. Introduction

The practice of parties agreeing on the procedures that will govern the resolution of their dispute is an inherent characteristic of various private mechanisms for dispute resolution, such as arbitration and mediation. In these processes, not only do the parties set the procedures that will apply to their dispute, but they also choose their own judge, set out the rules of evidence, and agree on the substantive applicable law.[1] By having considerable freedom to fashion the way that their dispute will be resolved, contracting parties can realize benefits that enhance their welfare.

However, applying a similar idea of private parties designing procedural rules in public litigation seems intuitively problematic. The private–public tension that parties' procedural rulemaking creates raises normative questions about the contours of parties' autonomy within adjudication: Should parties be allowed to depart from publicly created rules of procedure designed to guarantee procedural justice and a fair, efficient resolution of disputes? Should there be any limits to their freedom in customizing procedures? What criteria should inform the enforcement of private agreements setting procedures in public courts?

These questions concern procedural agreements made at two different points in time: before and after the dispute arises. After the dispute arises and a claim is filed, litigants may agree on procedures that would govern in the course of litigation. These agreements are a product of the adversary model of litigation, which is characterized by litigants' control over the way their dispute is adjudicated.[2] Since the litigants, and not the court, can best

* Lecturer, Radzyner School of Law, Interdisciplinary Center, Herzliya.

** Professor, Radzyner School of Law, Interdisciplinary Center, Herzliya.

1. *See e.g.*, Christopher R. Drahozal & Peter B. Rutledge, *Contract and Procedure*, 94 MARQ. L. REV. 1103, 1115–19 (2011) (empirically showing that contracting parties enter elaborate arbitration agreements).

2. *See, e.g.*, DAVID LUBAN, LAWYERS AND JUSTICE: AN ETHICAL STUDY 49–67 (1988); Lon L. Fuller, *The Forms and Limits of Adjudication*, 92 HARV. L. REV. 353, 364 (1978) ("[T]he distinguishing characteristic of adjudication lies in the fact that it confers on the affected party a peculiar form of participation in the decision, that of presenting proofs and reasoned arguments for a decision in his favor."); Michael L. Moffitt, *Customized Litigation: The Case for Making Civil Procedure Negotiable*, 75 GEO. WASH. L. REV. 461, 479–81 (2007) (arguing that procedural contracts promote procedural justice values by enhancing party control over the process); Ellen E. Sward, *Values, Ideology, and the Evolution of the Adversary System*, 64 IND. L.J. 301, 302 (1989)

represent their interests, allowing them to agree on the procedures that will apply during trial can reduce their costs, lower their risks, and guarantee a fair outcome. Indeed, the Federal Rules of Civil Procedure give litigants wide latitude to agree on various procedures.[3] For example, they may enter stipulations,[4] consent to waiver of service of process,[5] amend pleadings,[6] waive the right to a jury trial,[7] and agree on the extent of discovery proceedings[8] or on the taking of depositions.[9]

Before the dispute arises, parties may enter pre-dispute agreements that set out procedures that will apply to the resolution of future disputes within public adjudication. These agreements, which are mostly made as part of a contract stipulating the parties' substantive rights and obligations, may include various procedural matters such as the statute of limitations, interim measures, trial by jury, the scope of discovery, and the rules of evidence.

The concept of agreements that set procedures in adjudication has only recently begun to attract academic attention.[10] Robert G. Bone's *Party Rulemaking: Making Procedural Rules Through Party Choice*[11] is an outstanding contribution to the scholarship that seeks to define the contours of parties' procedural freedom within public courts. Bone examines the arguments for and against party rulemaking and clarifies the difficulty in

("The adversary system is characterized by party control of the investigation and presentation of evidence and argument, and by a passive decisionmaker who merely listens to both sides and renders a decision based on what she has heard.").

3. *See, e.g.*, STEPHEN C. YEAZELL, CIVIL PROCEDURE 138 (7th ed. 2008) ("One of the hallmarks of the U.S. law is the extent to which the rules of procedure are 'default' rules, rules that govern if the parties have not agreed to something else.").

4. *See* 73 AM. JUR. 2D *Stipulations* § 15 (2012).

5. FED. R. CIV. P. 4(d).

6. FED. R. CIV. P. 15.

7. FED. R. CIV. P. 39.

8. FED. R. CIV. P. 29.

9. FED. R. CIV. P. 29(a).

10. *See, e.g.*, Sarah Rudolph Cole, *Managerial Litigants? The Overlooked Problem of Party Autonomy in Dispute Resolution*, 51 HASTINGS L.J. 1199 (2000); Kevin E. Davis & Helen Hershkoff, *Contracting for Procedure*, 53 WM. & MARY L. REV. 507 (2011); Jaime Dodge, *The Limits of Procedural Private Ordering*, 97 VA. L. REV. 723 (2011); Drahozal & Rutledge, *supra* note 1; Daphna Kapeliuk & Alon Klement, *Contracting Around* Twombly, 60 DEPAUL L. REV. 1 (2010); Moffitt, *supra* note 2; Henry S. Noyes, *If You (Re)Build It, They Will Come: Contracts to Remake the Rules of Litigation in Arbitration's Image*, 30 HARV. J.L. & PUB. POL'Y 579 (2007); Judith Resnik, *Procedure as Contract*, 80 NOTRE DAME L. REV. 593 (2005); Robert J. Rhee, *Toward Procedural Optionality: Private Ordering of Public Adjudication*, 84 N.Y.U. L. REV. 514 (2009); Robert E. Scott & George E. Triantis, *Anticipating Litigation in Contract Design*, 115 YALE L.J. 814 (2006); Jean R. Sternlight, *Mandatory Binding Arbitration and the Demise of the Seventh Amendment Right to a Jury Trial*, 16 OHIO ST. J. ON DISP. RESOL. 669 (2001); David H. Taylor & Sara M. Cliffe, *Civil Procedure by Contract: A Convoluted Confluence of Private Contract and Public Procedure in Need of Congressional Control*, 35 U. RICH. L. REV. 1085 (2002); Elizabeth Thornburg, *Designer Trials*, 2006 J. DISP. RESOL. 181.

11. Robert G. Bone, *Party Rulemaking: Making Procedural Rules Through Party Choice*, 90 TEXAS L. REV. 1329 (2012).

assessing its limits from both a utilitarian and a right-based perspective.[12] He then concludes that party rulemaking should be limited in three distinct situations: when parties mutually agree to exclude a third party whose legal rights might be affected; when the procedural agreement is one-sided; and when the agreement restricts private enforcement of substantive law or conflicts with a proper consideration of civil rights claims.[13] For other situations Bone suggests that if party autonomy in fashioning procedure is to be limited, it must be because it risks threatening the normative legitimacy of public adjudication.[14] Consequently, Bone identifies adjudication's core characteristic as its commitment to a distinctive method of reasoning. He explains that "because the reasoning process is central to adjudication, we should focus on those procedural rules that have a strong effect on how that process is conducted."[15] Thus, procedural agreements that imperil the "procedures that frame, guide, or incentivize this reasoning process" should not be enforced.[16]

This Essay, in response to *Party Rulemaking*, focuses on an important aspect that Bone's rigorous analysis tends to undermine—the divergence between pre-dispute (*ex ante*) and post-dispute (*ex post*) procedural agreements.[17] This focus is necessary in order to understand the different private and public implications of these agreements. In an earlier manuscript, *Contractualizing Procedure*,[18] we focused on the private implications of party rulemaking. We showed how the different timing of procedural agreements—*ex ante* or *ex post*—affects the various advantages that parties can gain. In this Essay we push our analysis further, by focusing on the public implications of party rulemaking.[19]

The Essay proceeds as follows: Part II defines the situations in which *ex ante* and *ex post* party rulemaking change the litigation game that procedural rules define. Our main observation is that any mutual commitment to constrain, extend, or substitute the set of permissible actions defined by a

12. *Id.* at 1380–84.

13. *Id.* at 1383–84, 1397–98.

14. *Id.* at 1378–80.

15. *Id.* at 1391.

16. *Id.* at 1337.

17. *Id.* at 1340–41 (arguing that the distinction between *ex ante* and *ex post* is not that sharp). For a critical analysis of an *ex ante* perspective of procedural rules, see generally Robert G. Bone, *Agreeing to Fair Process: The Problem with Contractarian Theories of Procedural Fairness*, 83 B.U. L. REV. 485 (2003) [hereinafter Bone, *Agreeing to Fair Process*].

18. Daphna Kapeliuk & Alon Klement, Contractualizing Procedure (Dec. 31, 2008) (unpublished manuscript), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1323056.

19. Just like in *Contractualizing Procedure*, *id.*, our analysis here concerns procedural agreements bargained for by sophisticated commercial parties who knowingly consented to modify procedures. We do not discuss procedural agreements made in the case of an imbalance of bargaining power, agreements that are unenforceable under contract law, agreements that impair third parties, or agreements that conflict with substantive law.

procedural rule modifies the procedural rule and hence the litigation game. Part III discusses the different private implications of *ex ante* and *ex post* modified procedures. We describe the benefits that parties can gain by agreeing to modify procedures *ex ante* as compared to *ex post*. Part IV analyzes the public implications of agreements that modify procedure. We suggest that the public costs and benefits of *ex ante* agreements should not be evaluated from an *ex post* perspective only but also from an *ex ante* perspective. Our main argument is that since litigation is only one of many possible contingencies that could have taken place once the parties agreed to modify procedure, the court should discount the public costs it observes, *ex post*, by the probability that these costs would materialize, *ex ante*. Part V examines how the analysis of modified procedures in public adjudication should be informed by the alternative of private arbitration. We argue that since adjudication and arbitration are private substitutes, and since adjudication of contractual disputes creates not only negative externalities but also positive public goods, the *relative* flexibility of modifying procedures in these two alternatives should be taken into account. Part VI concludes.

II. Changing the Litigation Game

Party Rulemaking makes an important distinction between *actions* and *rules*. It explains that "[t]he general procedural rules define the set of permissible actions, and lawyers for the parties choose actions from the permissible set."[20] Accordingly, general procedural rules are based on the premise of a *constrained choice* of actions by litigants; they may choose any action within the set of actions defined by the procedural rule, yet they are limited to choose only from this set.[21]

When referring to the choice of an action from the set of actions defined by the rule in the course of litigation, Bone makes an interesting analogy to a "game tree" in game theory. "Procedural rules create the rules of the litigation game. They identify those who can be parties to litigation (the players of the game), the stages at which choices can be made, and the actions available to each party at each stage."[22] He thus refers to what game theorists call an "extensive form game."[23] The order of moves in an

20. Bone, *supra* note 11, at 1338. Bone refers later to a possible choice among different sets of rules through choice of forum. *Id.* at 1338–39. But, as he writes, "[l]ike choice of action in the previous paragraph, however, choosing rules by choosing a forum does not place party choice in conflict with what the chosen court would otherwise have applied. It merely selects among different sets of official rules." *Id.*

21. Most rules also allow for no action.

22. Bone, *supra* note 11, at 1338 n.37.

23. *See, e.g.*, DOUGLAS G. BAIRD ET AL., GAME THEORY AND THE LAW 307 (1994) (defining an *extensive form game* as containing the following elements: (1) the *players* in the game; (2) when each player can take an action; (3) what choices are available to a player when that player can act; (4) what each player knows about actions the other player has taken when that player decides what

extensive form game is represented in a game tree. Each tree has nodes.[24] In each node a player has to choose from a set of actions—branches. At the end of each final branch there are terminal nodes, which set out the payoffs that the players receive.[25] Bone's characterization of a game tree relates to the assignment of the nodes and actions as the rules of procedure. In addition, procedural rules, such as burden of proof and fee shifting, also affect the payoffs to the litigants, in addition to substantive law.[26] These rules are not directed at the parties and therefore cannot be defined as part of their "set of actions"; yet they are obviously structuring the litigation and its outcomes. Finally, there are other rules which do not form part of the very abstract game form, such as rules that relate to the way the courts have to act and reason. These rules, too, impact the process and the outcome of litigation. Procedural rules, thus, define a *litigation game.*

Bone offers a typology of the various possibilities for the parties to make procedure.[27] Types I and IV are unilateral actions taken in the course of litigation.[28] Types II and III are procedures agreed upon before the dispute arises but also afterwards. *Party Rulemaking* focuses on the last two types. In Type II procedures, parties "commit in advance to the same actions they could choose strategically during litigation."[29] In Type III procedures, parties consent "to a general rule different from the official rule that would otherwise apply."[30] So, the distinction between Type II and Type III party-made procedures is based on whether the action could be taken during litigation (Type II) or not (Type III). For example, according to Bone, an agreement to lengthen the statute of limitations is a Type II agreement, while an agreement to shorten it is a Type III rulemaking.[31] He explains that an extension of the statute of limitations could be achieved noncooperatively through waiver, whereas a shorter period is not possible to achieve during litigation.[32]

Bone's typology, although appealing, does not take into account whether and how these choices affect the litigation game. According to Bone, a Type II rulemaking is a "simple rule"[33] which "does not alter the

---

actions to take; and the *payoffs* to each player that result from each possible combination of actions); *see also id.* at 50–57 (analyzing the *extensive form game*). For a more formal definition of extensive form games, see DREW FUDENBERG & JEAN TIROLE, GAME THEORY 77–83 (1991).

24. BAIRD ET AL., *supra* note 23, at 311 (noting that a node is "the fundamental building block of the *extensive form game*").

25. *Id.* at 50.

26. Bone, *supra* note 11, at 1338 n.37 (noting that substantive and procedural law define litigation payoffs).

27. *Id.* at 1339–40.

28. *Id.* at 1339.

29. *Id.* at 1331, 1339–40.

30. *Id.* at 1332.

31. *Id.* at 1348.

32. *Id.*

33. *Id.* at 1338.

general rules that would otherwise apply; instead, it just moves to an earlier point in time a choice that the general rules allow parties to make later on."[34] However, the question should not be whether the chosen action is part of the original set of actions but whether the agreement *changes the set* of actions, thus imposing different constraints on parties' choice. Any such change alters the litigation game and consequently affects the range of possible contingencies that might emerge from it. It is when the litigation game changes that normative issues are at stake.

Moreover, Bone's typology undermines a key element necessary for assessing the limits of party autonomy to fashion procedure—whether the parties' choice is made before or after the dispute arises. As he writes: "Type II and Type III rulemaking can take place at any time before an action or rule is implemented. This includes during the course of litigation as well as before a lawsuit arises."[35] However, since any change in the litigation game changes the potential contingencies that could take place, the importance of the difference in timing of private rulemaking becomes apparent. When the change is made before the dispute, the range of contingencies that is affected is broader than when it is made after the dispute arises.

As we now show, both an agreement to commit to a particular permissible action within the original set and an agreement that allows for an action which is outside the set modify the otherwise applicable rule[36] and, thus, the litigation game. This change has strategic implications over the parties' behavior and over the outcome of the dispute. Moreover, these implications depend on the timing of the change made—before or after the dispute arises.

When the parties are engaged in litigation, they choose actions that best serve their interests. By selecting an action from the set defined by the procedural rule, a litigant also chooses *not* to take any of the other possible actions in the same set. Such choices are made unilaterally, as part of the strategic litigation game. For example, when a claimant waives her right to a jury trial[37] or decides to serve a limited number of interrogatories on the defendant,[38] or when a defendant waives service[39] or his statute of limitations defense,[40] they do not modify the procedural rule. They merely act according to the rule. In these cases the litigation game does not change.

When the parties agree on an action which is outside the set of actions that the rule defines, they modify the rule and, therefore, change the litigation

34. *Id.*
35. *Id.* at 1340.
36. Bone considers this possibility only in Type III rulemaking. *Id.*
37. FED. R. CIV. P. 38(d).
38. FED. R. CIV. P. 34 limits the number of interrogatories to twenty-five.
39. FED. R. CIV. P. 4(d).
40. FED. R. CIV. P. 8(c).

game. This is what Bone defines as Type III rulemaking.[41] However, an agreement to take a specific action which is part of the permissible set of actions according to the rule—a Type II rulemaking according to Bone—may also alter the rule. Such an agreement changes the set of actions that the rule defines and consequently the litigation game.

Contracting parties may modify procedural rules after the dispute arises, but also before it materializes—at the time of contracting. At the post-dispute stage, when the parties' agreement modifies the original set of permissible actions by restricting it to a subset of these actions (possibly only one of them) it changes the rule that would otherwise apply and, thus, modifies the litigation game. Likewise, when parties commit to an action which is outside the original set of actions, and therefore could not be taken otherwise, they modify the applicable rule. Once the modified procedure is agreed upon, the parties' strategic behavior changes, as each of them recognizes the new set of actions from which his opponent may choose. This situation is different from a unilateral choice of a specific action or a waiver of a possible objection to a specific action, which do not amount to a change of the general rule.

The following example clarifies the distinction between an *ex post* implementation of a procedural rule and an *ex post* change of a rule. Suppose that the rule provides that a claimant must bring suit no later than four years after the day the cause of action accrues and that the statute of limitations defense can be waived by the defendant. When the claimant brings suit after the prescribed period and the defendant waives his defense, the litigants do not modify the rule. They act according to the permissible set of actions that it defines. Thus, the litigation game is not altered. However, when the parties agree after the dispute arises to extend the statute of limitations, they modify the rule, by limiting the original set of possible actions available to the defendant. Since the claimant knows that if she brings suit the defendant *will not be able to object*, the modified procedure affects her decision whether to pursue litigation or to settle, and thus implicates the whole course of settlement negotiations and litigation between her and the defendant.[42]

Parties may also agree to expand the possible set of actions *ex ante*, ahead of the dispute. This would clearly amount to changing the applicable procedural rule. However, the rule would also change if they *commit ex ante* to take a particular action from the original set of actions defined by the rule or to avoid taking a specific action which could have been taken otherwise. According to Bone, this prior commitment is a simple rule which "does not alter the general rules that would otherwise apply; instead, it just moves to an earlier point in time a choice that the general rules allow parties to make later

41. Bone, *supra* note 11, at 1340.

42. Likewise, when the parties agree *ex post* to shorten the statute of limitation, they change the original set of permissible actions, and hence, they modify the rule.

on."[43] However, the parties' agreement to an action limits their future ability to choose other permissible actions from the original set. A prior commitment to a particular action from a set of available actions changes the original set of actions. As such, it necessarily modifies the litigation game.

For example, when the parties agree ahead of the dispute to forgo their Seventh Amendment right to trial by jury, to limit discovery, to limit the joinder of additional parties to a future claim, or to shorten or lengthen the statute of limitation,[44] they modify the otherwise applicable rule. From the parties' *ex ante* and *ex post* perspectives the only rule that will govern their dispute is the one that embodies their agreement. If their commitment did not amount to a modified rule, the parties would have no reason to commit to it ahead of the dispute. They would wait until the dispute arises and then choose any of the permissible actions within the original set.[45] Therefore, their pre-dispute commitment is an indication that the litigation game has changed.

To summarize, an *ex post* unilateral choice of action from a set of permissible actions is an implementation of the rule which does not change the litigation game. An *ex ante* or an *ex post* commitment to constrain, to extend, or to add to the original set of permissible actions defined by the rule modifies the rule, and thus changes the litigation game that would have taken place absent such commitment. Using Bone's typology, both Type II and Type III party rulemakings change the litigation game.

Having clarified when a procedural rule is modified, the next step in our analysis is to recognize that the same modified rules have different implications when they are agreed upon *ex ante*—before the dispute—than when they are agreed upon *ex post*—after it arises. Since commitments to modify procedure are made by private choice and since the parties' interests before the dispute are different than their interests after it arises, the same commitment must have different implications for the parties if it is made before the dispute than if it was made after it arises. Moreover, the range of potential future contingencies that the parties contemplate is fundamentally different, depending on the timing of the modification. These differences may not be observed in the specific case where the court is called upon to enforce the modified rule but only through an examination of the selection of

43. Bone, *supra* note 11, at 1338.

44. Bone's analysis of these two commitments show how his criteria fail to distinguish between cases that change the litigation game and cases that do not. According to Bone, "[a]n agreement to lengthen the statute of limitations is a form of Type II rulemaking. The statute of limitations is a waivable defense, so the same result could be achieved noncooperatively by waiver after the plaintiff files." *Id.* at 1348 (footnote omitted). However, an agreement to shorten a statute of limitations is, according to Bone, a Type III rulemaking. *Id.* Our analysis shows that both agreements change the rule, as the otherwise permissible set of actions defined by the rule are not applicable after an agreement is made. In both cases the parties modified the procedure, and hence, modified the litigation game.

45. In fact, any changes of the rule that are expected to be implemented *ex post* will not be agreed to *ex ante*.

cases and procedures that was induced by the parties' private choice. Just as settlements induce selection of cases for trial,[46] so do modified rules. Such selection depends, among other factors, on the time of modification.

The distinction between an *ex ante* and an *ex post* agreement to modify a procedural rule is, therefore, crucial to the analysis of its implications. In the following Parts we compare the implications of *ex post* and *ex ante* modified rules from two perspectives: private and public.

III. The Distinction Between *Ex Post* and *Ex Ante* Modified Procedures: The Private Perspective

Contractualizing Procedure[47] focused on the different private implications of *ex post* and *ex ante* agreements modifying procedural rules.[48] The paper explained how the different timing of these agreements affects the range of benefits that parties can gain from them. The dividing point—the dispute—is significant because it changes the interest structure of the parties. As we showed, *ex post*, the litigants can achieve limited benefits from procedural cooperation, whereas *ex ante*, they can gain substantial advantages from modified rules both before and after the dispute emerges.[49] *Contractualizing Procedure* described three major advantages that contracting parties can achieve through *ex ante* modified rules and that cannot be obtained through *ex post* cooperation.

First, modified rules can reduce strategic and opportunistic behavior and litigation costs should a dispute arise.[50] It is often the case that during trial, litigants abuse various procedural mechanisms, such as discovery and provisional remedies, to impose excessive costs on their counterparts. The parties are locked in strategic Prisoner's Dilemma situations, where each litigant tries to gain advantage over her counterpart by not cooperating. Their failure to cooperate might therefore amount to a mutual loss. At the pre-dispute stage the parties hold incomplete information about their future position in the post-dispute stage.[51] They do not necessarily know who will

---

46. George L. Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J. LEGAL STUD. 1 (1984).

47. Kapeliuk & Klement, *supra* note 18.

48. Here too, the paper assumed that parties to a commercial contract consider altering procedural rules and then, knowingly and voluntarily, enter procedural agreements. We ruled out procedural arrangements that are unenforceable under contract law and ruled out all possible claims of mistake, misrepresentation, fraud, oppression, duress, undue influence, or some other claim of unconscionability.

49. The distinction between the effects of *ex post* and *ex ante* agreements was first presented by Steven Shavell. Steven Shavell, *Alternative Dispute Resolution: An Economic Analysis*, 24 J. LEGAL STUD. 1 (1995). It was then further developed by Bruce Hay. Bruce L. Hay, *Procedural Justice—Ex Ante vs. Ex Post*, 44 UCLA L. REV. 1803, 1803–04 (1997) (arguing for an *ex ante* approach for evaluating the fairness of alternative dispute resolution mechanisms).

50. Kapeliuk & Klement, *supra* note 18, at 16–19.

51. Hay, *supra* note 49, at 1828–39 (describing the difference between *ex ante* and *ex post* perspectives when information differs). *But see* Bone, *Agreeing to Fair Process*, *supra* note 17, at

assume the role of plaintiff and who will be the defendant should a dispute arise.[52] This state of *common* uncertainty enables the parties to modify procedures, to avoid abuse of process, and to save costs, as long as their agreement does not adversely affect their incentives to perform their contractual obligations and comply with substantive law. Additionally, since prior to the dispute the parties are eager to cooperate, they could modify procedural rules in exchange for payments transferred between them.[53]

Second, modified procedural rules can shape the parties' *ex ante* substantive and procedural behavior. On the substantive level, modified procedures may engender incentive effects on the parties' behavior in performing their contractual obligations and complying with substantive law. On the procedural level, *ex ante* modified procedures can affect the parties' decision to engage in a dispute, to bring suit, to invest in litigation, and to consider the possibility of a settlement.[54]

The third advantage concerns the parties' ability to increase their welfare by creating information revelation mechanisms. We showed how uninformed parties can make use of modified procedures to screen and sort among potential partners based on their private information about their propensity to perform their contractual obligations and about their future litigation behavior, such as their propensity to use the legal system should a dispute arise, and the likelihood of using procedural mechanisms to impose costs on their counterparts.[55]

Ex ante, the interests of the parties are aligned. They can realize a mutual joint surplus by agreeing to modified procedural rules that would best accommodate their specific circumstances should a dispute arise and a claim is filed. *Ex post*, the litigants are engaged in a strategic game and are mainly concerned about distributional issues. They will consent to a particular action from a set of permissible actions or to an action outside the set, to the extent that the agreement reduces their costs, lowers their risks, and does not adversely affect the expected outcome. However, when their interests are opposed, cooperation might be difficult to achieve. This situation can be described in game theory as a "zero-sum game." In this game the gain of one

526–29 (criticizing what he defines the "*ex ante* argument" as an argument that imposes "substantial restrictions on the parties' knowledge" and arguing that the assumptions underlying the argument are too strong).

52. For a criticism on this assumption, *see* Bone, *supra* note 11, at 1341 ("[T]he informational differences between *ex ante* and *ex post* are not as stark as some commentators assume.").

53. Kapeliuk & Klement, *supra* note 18, at 18. Bone argues that side payments are also possible during litigation; however, as he rightly acknowledges, "there is more room for making side payments *ex ante*." Bone, *supra* note 11, at 1341.

54. Kapeliuk & Klement, *supra* note 18, at 19–23.

55. *Id.* at 23–25. Another benefit explored by Scott and Triantis is the ability of parties to shift costs between the pre-dispute and the post-dispute stages. *See* Scott & Triantis, *supra* note 10.

player necessarily implies the loss of the other player.[56] Therefore, in situations where the choice benefits a litigant at the expense of the other, an agreement is unlikely.[57] Among the reasons for negotiation failure are rational[58] and behavioral[59] reasons. Most commonly, the only way to overcome such barriers is through an agreement on an overall settlement.

Bone claims that *Contractualizing Procedure* exaggerates the difference between the likelihood of cooperation *ex ante* and *ex post*.[60] Whether this is true or not is a matter for empirical examination. Yet, even if the difference is not that stark, this does not undermine the fundamental distinction between *ex ante* and *ex post* agreements to modify procedure, as far as the possible private benefits to be reaped from such agreements are concerned. *Contractualizing Procedure*'s main contribution was to highlight this distinction.

IV. The Distinction Between *Ex Post* and *Ex Ante* Modified Procedures: The Public Perspective

We now push the analysis one step further. We discuss how the *ex ante* perspective should affect the public—as distinguished from the private—analysis of costs and benefits of customized procedures. As we explain, an *ex ante* approach allows evaluation of public benefits and costs that are not manifested, *ex post*, in the case before the court. Furthermore, we show that the public costs of pre-dispute modified rules, however large they are, should be discounted by the probability that they would materialize. Such discounting is absent from scholarly literature analyzing *ex ante* procedural agreements,[61] including *Party Rulemaking*.

---

56. *See, e.g.*, BAIRD ET AL., *supra* note 23, at 317 (defining a zero-sum game as "[a] game in which the increase in the *payoff* to one player from one combination of *strategies* being played relative to another is associated with a corresponding decrease in the *payoff* to the other").

57. *See* Kapeliuk & Klement, *supra* note 10 (analyzing this problem in the case of pleading standards).

58. Under the assumption of rationality, the main reason for negotiation failure is information asymmetry between the parties. *See, e.g.*, Lucian Arye Bebchuk, *Litigation and Settlement Under Imperfect Information*, 15 RAND J. ECON. 404 (1984).

59. These include reactive devaluation (negotiating parties discount the value of settlement offers only because they were made by their counterparts), endowment effects (negotiating parties find it difficult to forgo part of their perceived rights and value those rights more than if they did not perceive to own them), over-optimism (each party is overly optimistic about her trial prospects), and framing effects. *See generally* Daniel Kahneman & Amos Tversky, *Conflict Resolution: A Cognitive Perspective*, *in* BARRIERS TO CONFLICT RESOLUTION 45 (Kenneth J. Arrow et al. eds., 1995); Russell Korobkin, *Psychological Impediments to Mediation Success: Theory and Practice*, 21 OHIO ST. J. ON DISP. RESOL. 281 (2006).

60. Bone, *Agreeing to Fair Process*, *supra* note 17, at 491 (noting that "a procedure is fair if all parties would have agreed to the procedure had they been able to contract for it in advance of ("ex ante") their dispute").

61. For the literature on procedural rulemaking, *see* Kapeliuk & Klement, *supra* note 10.

In discussing these issues we define costs and benefits in broad, not necessarily economic, terms.[62] The costs are thus not only direct (opportunity costs of judiciary, jury, and administrative personnel, rent, and other resources) and indirect (such as congestion and delay costs imposed on third parties who wait for their turn to litigate and error costs), but also other, more abstract, costs. For example, an agreement that adversely impacts judicial legitimacy is considered as a public cost.[63] True, this broad perspective limits the possibility of conducting a comprehensive welfare analysis, as it is done by economists.[64] Absent a common denominator, how should one weigh, for example, adverse legitimacy effects against savings in litigation costs and improved contractual incentives? Still, this approach enables realization of what is at stake when analyzing the normative boundaries of parties' freedom to customize procedures.

A. The Public Implications of Procedural Modifications

Scholars considering the public implications of modified procedures focus on the *ex post* effects of these procedures. They examine how the enforcement of the parties' agreement would affect the direct and indirect costs of the judicial system as well as the court's legitimacy.[65] Bone's analysis of judicial legitimacy is no exception in this respect.

An *ex post* approach is appropriate if the enforcement of a post-dispute procedural agreement is at stake, and if such agreement could not be anticipated by the parties before the dispute. This is usually the case with *ex post* procedural modifications, in view of the conditions that must be satisfied for them to materialize, as explained in the previous Part.

However, this approach is problematic when examining a pre-dispute agreement to customize procedure, since it overlooks the broad effects of the parties' agreement, as they are evaluated from an *ex ante* perspective. Instead of analyzing how the parties' agreement has altered the litigation game, an *ex post* approach examines how one possible outcome has changed. Thus, an *ex post* approach mistakes one branch of the tree for the whole tree. For some customized procedures such omissions are inconsequential. However, often this would not be the case.

When evaluating the impact of a pre-dispute procedural commitment on institutional values such as judicial integrity and legitimacy, courts should be aware that the outcome they observe is only one of many possible contingencies that could have materialized. The modified rule has transformed the parties' relationship from the time they had agreed on it. It has affected their behavior in performing their contractual obligations, the

62. For Bone's analysis of the costs of party rulemaking, *see* Bone, *supra* note 11, at 1372–80.

63. *Id.* at 1378–80.

64. As Bone rightly suggests, *id.* at 1381.

65. Refer to Bone's detailed analysis, *id.* at 1374–80.

probability that a dispute would arise, and their litigation behavior. All these effects have public implications that go beyond the parties and that have to be considered when enforcement of the parties' agreement is at stake. Focusing on one contingency that has materialized misses the full range of public implications.

Suppose, initially, that an agreement modifying a procedural rule affects the direct and indirect costs of the judicial system but that it does not impact the court's institutional legitimacy. Suppose also that the modified rule increases the demand on the court's time and costs. For the purpose of clarification, we take an example of an agreement to allow broad discovery.[66]

When the parties opt for such an agreement after the dispute arises, and such an agreement could not be anticipated in advance, the only implications of the agreement may be those that are observable.[67] When considering whether to enforce the agreement, the court may find that the agreement increases the overall public direct and indirect costs, and therefore decline to enforce it.

However, an agreement made before the dispute arises involves conflicting effects on the overall costs of litigation, case backlog, and case delay. An increase of these costs in a specific case would also enhance litigants' incentives to settle and discourage their desire to litigate. Thus, a modified rule may increase the costs of a specific procedure but at the same time decrease the overall litigation costs since less cases would be filed. And of those filed, more cases would settle. The overall change in litigation costs would be the sum of these conflicting effects, which may be either positive or negative. A court focusing on the direct costs of broader discovery from an *ex post* perspective, would miss the pre-dispute agreement's potential for discouraging litigation and encouraging settlement, both effects reducing public litigation costs and delay.

Similar reasoning holds for broader institutional implications. Suppose that the criterion for evaluating the enforcement of a modified procedure is the one that Bone suggests: namely, that "the core element of adjudication is its distinctive mode of principled reasoning."[68] The effect of a pre-dispute agreement to modify procedure must then be evaluated by weighing its aggregate effect over this core element of adjudication. Aggregation must be conducted over all possible contingencies, viewed and weighted by their likelihood from an *ex ante* perspective.

Take, for example, a motion by the plaintiff for the discovery of electronically stored information against the objection of the defendant who,

---

66. FED. R. CIV. P. 29.

67. If, prior to the dispute, the parties expect such agreement to take place, then this expectation would clearly affect their pre-dispute behavior. But then, the expected agreement is part of the litigation game, and it does not change it. For the analysis of procedural rules from an *ex ante* perspective, see Hay, *supra* note 49.

68. Bone, *supra* note 11, at 1390.

in turn, relies on a pre-dispute agreement not to make any such discovery in case of litigation. As we explained above, the agreement is a modified rule since it limits the otherwise available options that the rule defines.

The court may consider this motion on its specific merits and examine how the documents might affect the process and its outcome. It may find that absent such discovery the plaintiff would be unable to uncover certain facts, which are essential for a proper and accurate decision-making process. Given the central role of discovery in the Federal Rules of Civil Procedure,[69] the court may find this agreement to be detrimental to its ability to apply principled reasoning and to decide the case correctly, on its *true* merits.[70]

But the implications of this agreement over the court's ability to apply principled reasoning and reach an accurate decision may go way beyond the merits of the specific outcome in this specific case. The distinction between a pre-dispute and a post-dispute agreement is crucial here. Parties may make a pre-dispute agreement denying e-discovery for various reasons: to save costs, to induce each party to keep better record of future events, to limit opportunistic behavior in litigation, or to encourage early settlements. Any of these private goals has also public consequences. From an *ex ante* perspective this agreement combines possible contingencies that would be most conducive to a fair and reasoned adjudicative process, with the possibility of an eventuality in which the lack of e-discovery would critically affect the judicial process and its outcome. These types of contingencies are characteristic of a pre-dispute procedural agreement and are by and large absent when a similar agreement is made after the dispute arises.

Ex post, the litigants establish expectations about the possible outcome of the case and the contribution of discovered documents to the outcome. Both would therefore agree to limit discovery or forgo it altogether only if the effect on the expected outcome is not significant enough to overcome their joint savings in costs and delay. If the outcome is expected to be significantly affected by undiscovered documents, then one of the parties, at least, would refuse to forgo discovery. However, these conditions can hardly be anticipated before the dispute.

A court considering whether to enforce such an agreement should not find it very problematic to decide on the matter. First, the specific documents in the specific case are all that it should consider. There are no other "branches" of the litigation game that were affected as the agreement was made down the road after litigation embarked. Second, the court may presume that the documents are not indispensable for delivering a correct decision. Otherwise the litigants would have been unlikely to agree on it. In

69. 8 CHARLES ALAN WRIGHT ET AL., FEDERAL PRACTICE AND PROCEDURE § 2001 (3d ed. 2010).

70. *But see* Bone, *supra* note 11, at 1391 ("[T]here is no way to identify particular procedures that strongly affect the reasoning process because all procedures have the capacity to do so . . . .").

fact, for this same reason, the court would be unlikely to be called upon to revoke such an agreement, as none of the litigants would so require.

The same modification, agreed to behind the "veil of incomplete information" before the dispute, has altogether different implications. By the time of the agreement the parties do not know which documents would be affected by such an agreement. They may contemplate future contingencies in which a document would be of prohibitive value in litigation. Nevertheless, in view of the private disadvantages of future e-discovery, they may still constrain it.[71] As a result, such contingencies might materialize, potentially leading to a significant divergence between the expected judicial outcome and the true merits of the case. Yet, just as the parties weigh possible future contingencies (including adverse ones) when contemplating a constraint on discovery, so should the court try to evaluate what were the public implications of such a constraint at the time of contracting.

Since this evaluation is done in retrospect, it is most difficult to implement for reasons we explicate below. Therefore, one cannot realistically advocate a fully blown cost–benefit analysis (those terms broadly defined) of the discovery constraint. Still, in view of the potential positive implications of such a constraint, including inducement of better process and outcome in many contingencies, a court may be more likely to tolerate a specific divergent outcome. Evaluating it against the full spectrum of possible contingencies, viewed before the dispute, should prove different than narrowly focusing on its direct and specific implications for the case before the court.

True, when the negative value attributed to a particular modified rule is high, it may be sufficient to render it unenforceable, irrespective of the likelihood of litigation taking place. For example, a modified rule that directly regulates the conduct and the decision-making process of the court might carry such a red flag over it that courts cannot tolerate enforcing it, rare as its realization might be.[72] This may be true independent of the modified rule's potential private and public merits, as they are viewed from a pre-dispute perspective.

However, there is a broad selection of modified procedures, which might impact the process and outcome of adjudication, even though these procedures are directed at the parties and not at the court. When considering whether to enforce any such modified rule *ex post* the court ought to understand the modification's broad implications, not only the effects in the specific case before it. What might seem problematic in a specific case

71. For the benefits that contracting parties can gain from such agreement, see Kapeliuk & Klement, *supra* note 18, at 16–23.

72. Bone considers these rules as "rules defining the decision-making body and the decision protocol, including not only those directed to the judge but also to the jury." Bone, *supra* note 11, at 1393.

might prove valuable, based on the same normative criterion, if viewed from an *ex ante* perspective.

We do not make here any claim as to what should be the criterion for deciding which procedural modifications are to be enforced. Nor do we examine which virtues are fundamental to the legitimacy of adjudication so that interfering with them should not be tolerated. Rather, we highlight the significance of the *timing* of modification for the analysis of the modification's public implications. An examination of a modified rule, without reference to whether it was agreed before or after the dispute, risks missing its potential implications, positive and negative, from a public perspective.

B. The Inherent Difficulties of Analyzing Public Ex Ante *Costs and Benefits in Retrospect*

As Bone rightly explains, it is difficult for courts to analyze, *ex post*, the overall costs and benefits of pre-dispute procedural modifications.[73] This is true even if the analysis is restricted to purely economic costs and benefits. Broadening the perspective to the institutional legitimacy of the judicial system further complicates such analysis.

There are various reasons for this difficulty. First, courts observe *ex post* only one possibility which has materialized out of many potential contingencies. They are called to decide a specific dispute which has turned into a specific litigation following a specific course of action by the litigants. Courts do not see all other possibilities which did not come true either by chance or because of a deliberate decision by the parties. All contingencies where the parties have satisfied their contractual obligations, or have decided not to embark on litigation even if those obligations were not satisfied, are not considered *ex post*. *Ex post*, courts can only speculate, absent any concrete evidence, what these other possibilities would have been. Hence, analyzing their consequential costs and benefits is problematic.

Second, courts would find it very hard to identify and measure all the pre-dispute benefits induced by a modified procedure. We described above the conflicting implications of any change in litigation costs over total costs and delay. Consideration of other advantages of modified procedures such as improvements in incentives to perform contractual obligations—and, consequently, over the total value of a contract—and information revelation benefits, induced by the choice of a specific procedure, is even more complicated.

Do these difficulties imply that courts should only examine the *ex post* effects of a modified procedural rule and ignore all other potential contingencies, consequently ignoring the timing of modification? We believe not. Pre-dispute analysis is indeed difficult to conduct in retrospect.

---

73. *Id.* at 1381 (describing it as "impractical").

Yet, focusing only on observable *ex post* costs and benefits would be overly restrictive and inadequate. Just as substantive contract law focuses on the time of contracting, so should procedural contracts be analyzed from the same perspective.[74]

We are, therefore, aware of the potential problems in analyzing modified procedures from an *ex ante* perspective. Although this is unfortunate, we are unable, at this stage, to provide any easy way out of these problems. We only suggest that courts be aware of these complications and, in particular, that they take heed of the timing of the agreement to modify a procedure when considering its public implications.

V. The Arbitration Alternative

Parties may agree to modify procedural rules both before and after the dispute arises. Yet, their choice is not limited to the specific rules that will govern the adjudication of their dispute but also to the forum that will rule on its merits. While public courts are the default forum for the resolution of disputes, parties may agree to opt for a private forum to settle their controversies. They may do so at the contracting stage or after the dispute arises.

Whether the parties opt for public adjudication or for private arbitration, they may wish to customize the procedure that will govern their dispute so as to best serve their interests. When choosing arbitration, they have almost unrestricted freedom to fashion the procedure. In contrast, when opting for public adjudication, their choices might be subject to court approval. This difference may affect the parties' choice between arbitration and adjudication.

In the previous Part we argued that the analysis of the public implications of procedural agreements should focus not only on their *ex post* public costs and benefits but also on their *ex ante* public effects. Our main claim was that since litigation is only one of many possible contingencies that could have taken place once the parties entered their agreement, a court should discount the public costs it observes, *ex post*, by the probability that they would materialize, *ex ante*. We acknowledged, however, that an evaluation of the public costs and of the probability of their occurrence is difficult to implement in retrospect as courts do not necessarily possess the necessary tools to quantify them. This obvious difficulty[75] leaves us with the

74. As Bone rightly suggests, "there is an alternative to case-specific evaluation with all its prediction problems. The formal rulemaking process can be used to create general rules regulating party rulemaking." *Id.* at 1384. However, this framework would possibly prove too general, since the value and costs of pre-dispute modified procedures would often depend on the specific context in which they were modified.

75. Or, as Bone suggests, impracticability. *Id.* at 1381 ("A judge in an individual case lacks the information and expertise to make highly complex predictions about case-specific benefits and costs.").

question of whether courts should, as a result, deny the enforcement of procedural arrangements and thereby encourage parties who wish to customize procedure to opt for private arbitration. In order to address this question we return to the concept of externalities.

As we explained,[76] litigation creates various types of positive and negative externalities—public costs and benefits—that are not considered by the parties when they commit to modify procedures. Thus, there is a divergence between the private and public interests, both concerning the mere desirability of adjudication[77] and the desirability of enforcing specific modified procedures. It is the negative externalities, and the resulting misalignment of the public and private interests, which render specific modifications suspect and justify *ex post* judicial intervention.

In contrast, contracting parties who opt for arbitration internalize all costs and benefits of their future dispute should it arise, including those of the arbitration mechanism itself. Hence, if contracting parties adopt a certain procedure for their future disputes it must be efficient, in that its overall costs are lower than its overall benefits.

Bone rightly criticizes what he calls "[t]he flawed argument from arbitration."[78] His main point is that one cannot make a simple comparison between adjudication and arbitration since they perform different functions and draw on different sources for their legitimacy.[79] In our terminology, adjudication creates externalities which are absent in arbitration. This distinction between the two processes implies that one may not derive normative conclusions about modified procedures in adjudication from observing similar procedures in arbitration.[80] Bone is also right in arguing that the mere flight from adjudication to arbitration is not, by itself, a justification to render adjudication more flexible.[81]

We argue that whether substitution of arbitration for litigation is problematic depends on the net value of externalities, positive and negative, created by adjudication. If this net value is negative, then all contractual disputes should be resolved in arbitration. Since parties internalize all costs and benefits in private arbitration, their agreement to arbitrate must necessarily be efficient. Thus, by having all contractual disputes decided by arbitration, one can guarantee efficiency, narrowly defined.[82]

---

76. *See supra* Part IV.

77. *See generally* Steven Shavell, *The Fundamental Divergence Between the Private and the Social Motive to Use the Legal System*, 26 J. LEGAL STUD. 575 (1997).

78. Bone, *supra* note 11, at 1354.

79. *Id.*

80. *Id.*

81. *Id.* at 1354–55.

82. *See* Bruce L. Hay et al., *Litigating BP's Contribution Claims in Publicly Subsidized Courts: Should Contracting Parties Pay Their Own Way?*, 64 VAND. L. REV. 1919, 1948–50 (2011) (suggesting the imposition of a user fee on commercial contractual litigation to internalize all its costs and induce an efficient choice between litigation and arbitration).

However, one must also bear in mind that litigation does not produce only negative externalities but also positive public goods[83] such as legal precedents and deterrence effects,[84] as well as public substantiation of its institutional legitimacy.[85] When parties agree to modify procedural rules within adjudication they do not internalize the negative externalities that their agreement produces. But they do not internalize the public benefits of adjudication either. If one is to maintain litigation of contractual disputes, at least for adjudication's positive externalities, one cannot dismiss the substitution effect between the two institutions out of hand.

When the choice of process is made after the dispute arises, the parties are locked into litigation as the default procedure. Thus, constraining the parties' ability to modify procedures is unlikely to induce them to agree to opt out of it. But, when the choice is made before the dispute arises, rendering litigation less flexible would drive more parties to opt out of it. Since this choice has public implications, they must be taken into account when considering the flexibility that contracting parties are allowed in structuring future procedures within public adjudication.

VI. Conclusion

This Essay argued that timing matters. It matters with respect to the point in time in which parties agree to modify procedure, which, in turn, affects the private and public implications of modified procedures. *Party Rulemaking* is an important contribution to the emerging scholarship on modified procedures. Bone's identification of the core of adjudication's normative legitimacy—its distinctive mode of reasoning—as the threshold for enforcing modified procedures is without doubt an important step in delineating the limits of party choice in customizing procedures. Yet, his focus on *ex post* public implications is too narrow.

Our main argument was that an *ex post* perspective on the public implications of modified procedures is appropriate when parties agree to modify procedures after the dispute arises. However, an *ex post* approach to the public implications of *ex ante* modified procedures misses an important factor—the fact that the outcome that the court observes is only one of many possible contingencies that could have materialized once the parties agreed to a modified procedure. The modified procedure affected the parties' behavior in performing their contractual obligation, the probability that a dispute

---

83. A public good has two related characteristics: its consumption by one person does not leave less for others, and it is nonexclusive, meaning that once the good is produced, no one can be denied of its consumption, even if they have not paid for it. *See, e.g.*, ROBERT COOTER & THOMAS ULEN, LAW AND ECONOMICS 42 (3d ed. 2000).

84. For the analysis of the private and public goods provided by adjudication, see William M. Landes & Richard A. Posner, *Adjudication as a Private Good*, 8 J. LEGAL STUD. 235 (1979); Moffitt, *supra* note 2, at 519 ("Courts perform important functions in society beyond dispute resolution. Courts articulate community norms.").

85. Moffitt, *supra* note 2, at 519.

would arise, and their propensity to litigate or settle. These effects are not limited to the private domain but have also public implications, which should be considered when the normative question is at stake. An *ex post* approach, which focuses only on the contingency that has materialized, ignores the full range of public implications of the modified procedure. Thus, an appropriate approach should discount the public costs of such procedure by the probability that they would materialize. We acknowledge that a quantification of the *ex ante* public effects of modified procedures is difficult. Yet, we maintain that this difficulty cannot justify disregarding these effects.

Notes

No Mere "Matter of Choice": The Harm of Accent Preferences and English-Only Rules*

Introduction

Native-born members of democracies—perhaps driven by fears of economic and cultural usurpation—have long resented and felt threatened by immigrants.[1] Even the United States, a country famously built by immigrants, has a history of hostility towards immigrants that stretches back centuries.[2] Despite that widespread opposition, the number of immigrants to the United States is increasing all the time.[3] Each year, hundreds of thousands of immigrants move to the United States[4] hoping to take advantage of the economic, societal, and educational opportunities this country has to offer. With more immigrants, more contact between native English speakers and those who speak English as a second language is inevitable. That contact has spurred a number of conflicts between native-born Americans and immigrants and their children.[5] This Note will focus on the conflicts caused by the differences in languages of those two groups, particularly the problems caused by accent preferences and English-only rules.

---

* I would like to express my sincere gratitude to the editors and staff of the *Texas Law Review* for their hard work in turning my paper into a publishable Note, Professor Cary Franklin for her great advice and guidance during the writing of this Note, my caring family for all of their encouragement throughout law school, and above all, my wonderful fiancée Tänne for her constant love and support.

1. Pamela Paxton & Anthony Mughan, *What's to Fear from Immigrants? Creating an Assimilationist Threat Scale*, 27 POL. PSYCHOL. 549, 549–50 (2006).

2. *Id.*

3. *See* Randall Monger & James Yankay, *U.S. Legal Permanent Residents: 2011*, U.S. DEP'T HOMELAND SEC. ANN. FLOW REP. 1, 1 fig.1 (2011) (finding that the number of people becoming legal permanent residents is increasing).

4. *Id.* at 4 tbl.3; *see* Michael Hoefer et al., *Estimates of the Unauthorized Immigrant Population Residing in the United States: January 2011*, U.S. DEP'T HOMELAND SEC. POPULATION ESTIMATES 1, 5 tbl.3 (2012) (stating that the average annual change in unauthorized immigrant population from 2000 to 2011 was 280,000).

5. I would like to note at the outset that, particularly in the context of English-only rules, much of this Note will deal with Latino national origin and the Spanish language. That focus is not deliberate, but is instead a function of Latinos being the largest immigrant group to the United States. Monger & Yankay, *supra* note 3, at 4 tbl.3; Hoefer et al., *supra* note 4, at 5 tbl.3. Other national origin groups certainly bring lawsuits against employers because of English-only rules. *See, e.g.*, Kania v. Archdiocese of Phila., 14 F. Supp. 2d 730, 731 (E.D. Pa. 1998) (denying the defendant's motion to dismiss a claim against an English-only rule brought by a Polish speaker); Edward M. Chen, Garcia v. Spun Steak Co.: *Speak-English-Only Rules and the Demise of Workplace Pluralism*, 1 ASIAN L.J. 155, 159 n.22 (1994) (listing challenges to English-only rules brought by Asian Pacific Islanders). I intend for this analysis to apply to all foreign national origin groups.

Many employers have enacted "English-only rules" that prohibit the speaking of any languages other than English at work. Other employers have passed over immigrants[6] for employment opportunities because of their accents. These employers generally feel that they are justified in taking these actions based on the needs of their businesses, but since the 1980s, immigrants have regularly filed lawsuits challenging these policies and decisions.[7] Title VII of the Civil Rights Act of 1964[8] protects immigrants and their descendants from discrimination on the basis of national origin.[9] Immigrants who have been denied jobs because of their foreign accents have sued employers, alleging national origin discrimination.[10] They have also used Title VII to sue employers who institute English-only rules, claiming that the negative effect the rules have on them as non-native English speakers amounts to national origin discrimination.[11] Despite the strong links between national origin, language, and accents, courts have been reluctant to rule for plaintiffs on these claims. Courts generally find either that there is not significant harm to the plaintiff, or that the employer has sufficient business reasons to justify its decision or policy.[12]

Part I of this Note lays out the legal framework for these national origin discrimination claims. In Part II, this Note tells the stories of several plaintiffs who challenged English-only rules and accent-based hiring decisions. Part III argues that many courts are doing a great disservice to the goals of Title VII in the way they treat these claims. Part IV explores the harm caused by English-only rules and accent preferences based on the link between language and accent and one's national origin. To many immigrants, language and accent are very much a part of who they are and are not as mutable as courts generally assume. Kenji Yoshino argues that much of the discrimination that goes unchecked today involves forcing minorities to hide, or "cover," traits linked to their minority status, which does serious harm to the identities of members of those groups.[13] By forcing employees to cover their accents and native languages, employers are attacking the national origin identities of those employees. Part V argues that after recognizing the severity of those attacks, courts should analyze claims against English-only rules and accent discrimination differently and scrutinize employers' business justifications more closely.

---

6. For the sake of convenience, I will often use the term "immigrants" to refer to the employees of non-U.S. national origin who are affected by these rules. However, as will be explained below, first-generation immigrants, naturalized citizens, and the children of those immigrants are all protected equally under Title VII's definition of "national origin."

7. *See infra* Part II.

8. Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e (2006).

9. 42 U.S.C. § 2000e-2(a).

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

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

12. *See infra* subpart III(B).

13. Kenji Yoshino, *Covering*, 111 YALE L.J. 769, 781 (2002).

I. The Framework of Lawsuits Against English-Only Rules and Accent Discrimination

Title VII provides that:

> It shall be an unlawful employment practice for an employer—
>
> (1) to fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's . . . national origin; or
>
> (2) to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's . . . national origin.[14]

National origin is not formally defined in Title VII,[15] and did not get as much attention as the other protected categories during the legislative debates about Title VII.[16] However, since the passage of Title VII, national origin has been interpreted broadly by the Equal Employment Opportunity Commission (EEOC), which issues interpretive guidelines on employment discrimination issues, and by the Supreme Court. In *Espinoza v. Farah Manufacturing Co.*,[17] the Supreme Court examined the "quite meager" portion of Title VII's legislative history that dealt with national origin.[18] The Court found that national origin "on its face refers to the country where a person was born, or, more broadly, the country from which his or her ancestors came."[19] The EEOC later issued its Guidelines on Discrimination Because of National Origin, which define national origin discrimination as "the denial of equal employment opportunity because of an individual's, or his or her ancestor's, place of origin."[20] Some commentators complain about the vagueness of this definition[21] and attempt to reformulate the language to better protect people from discrimination on the basis of their national origin groups.[22] However,

---

14. 42 U.S.C. § 2000e-2(a).

15. *Id.* § 2000e.

16. *See* Espinoza v. Farah Mfg. Co., 414 U.S. 86, 88–89 (1973) (interpreting the meaning of "national origin" and noting that "[t]he statute's legislative history [is] quite meager in this respect").

17. 414 U.S. 86 (1973).

18. *Id.* at 88–89.

19. *Id.* at 88.

20. 29 C.F.R. § 1606.1 (2012).

21. *See, e.g.*, Mark Colón, *Line Drawing, Code Switching, and Spanish as Second-Hand Smoke: English-Only Workplace Rules and Bilingual Employees*, 20 YALE L. & POL'Y REV. 227, 231–32 (2002) (finding that the legislative history of Title VII and the Supreme Court's definition of "national origin" "offer[] no guidance regarding employment practices aimed at underlying personal characteristics, including language, that are often closely associated with national origin").

22. *See* Juan F. Perea, *Ethnicity and Prejudice: Reevaluating "National Origin" Discrimination Under Title VII*, 35 WM. & MARY L. REV. 805, 810 (1994) (arguing for an amendment to Title VII to better protect "ethnic traits [and] ethnicity").

taken at its plain meaning, this definition should provide protection against language and accent discrimination.

Plaintiffs use Title VII to bring several types of claims against employers for accent preferences and English-only rules. Plaintiffs who allege accent discrimination are mostly limited to claiming disparate treatment. To succeed in a disparate treatment suit, a plaintiff must show that (1) he was a member of a protected class; (2) he was qualified for a position; (3) an adverse employment action was taken against him; and (4) that adverse employment action "occurred under circumstances giving rise to an inference of discrimination."[23] To rebut that showing, the employer must establish that it actually took the adverse action because of some "legitimate, nondiscriminatory reason."[24] For an employer to have an acceptable business reason for discriminating against a plaintiff because of his accent, it must show that the accent "interferes materially with job performance."[25] If the employer establishes a legitimate nondiscriminatory reason, the plaintiff can attempt to show that the stated reason was actually a pretext for a prohibited motivation.[26] The plaintiff can also claim disparate treatment under a mixed-motive framework, although his remedies could be limited.[27] Under this framework, the employee would have to show that even if the employer had legitimate reasons for taking an adverse action against the employee, the employee's protected trait was still impermissibly considered.[28]

Immigrants who file suit against employers because of their English-only rules can bring claims for disparate impact, hostile work environment, and possibly, as Part V will argue, systemic disparate treatment. In a disparate impact case, plaintiffs allege that an employer has a policy or practice that is facially neutral, but whose effect disproportionately burdens a protected class.[29] A plaintiff does not have to show that the employer had a discriminatory intent to prevail on a disparate impact claim.[30] Disparate impact claims against English-only rules generally fall under 42 U.S.C. § 2000e-2(a)(1), so a plaintiff must show that the rule's uneven burden

23. *See In re* Rodriguez, 487 F.3d 1001, 1008 (6th Cir. 2007) (adapting the original claim structure laid out in *McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 802 (1973)).

24. *McDonnell Douglas Corp.*, 411 U.S. at 802.

25. Fragante v. City of Honolulu, 888 F.2d 591, 596 (9th Cir. 1989).

26. *McDonnell Douglas Corp.*, 411 U.S. at 804.

27. *See* 42 U.S.C. § 2000e-5(g)(2)(B) (2006) (limiting plaintiffs to declaratory and injunctive relief but not allowing damages for mixed-motive claims in which the employer shows it would have "taken the same action in the absence of the impermissible motivating factor").

28. *See* Price Waterhouse v. Hopkins, 490 U.S. 228, 252 (1989) (laying out this framework for mixed-motive cases); *Fragante*, 888 F.2d at 598 (examining an accent-discrimination plaintiff's claim for signs of an employer's mixed motive).

29. Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 335–36 n.15 (1977).

30. *Id.*

affects the "terms, conditions, or privileges of employment."[31] The Supreme Court has held that "[t]he phrase 'terms, conditions, or privileges of employment' evinces a congressional intent 'to strike at the entire spectrum of disparate treatment of men and women' in employment,"[32] and that Title VII "covers more than '"terms" and "conditions" in the narrow contractual sense.'"[33] Courts have acknowledged that "policies or practices that impose significantly harsher burdens on a protected group than on the employee population in general may operate as barriers to equality in the workplace and . . . may be considered 'discriminatory.'"[34] If a plaintiff can show that a policy has that effect, the burden shifts to the employer to prove that its policy is "consistent with business necessity."[35] Finally, if an employer can establish business necessity, the plaintiff must prove that there is an "alternative employment practice" that can accomplish the employer's business goals with a less discriminatory effect.[36]

To bring a hostile work environment claim, a plaintiff must show that his "workplace is permeated with 'discriminatory intimidation, ridicule, and insult,' . . . that is 'sufficiently severe or pervasive to alter the conditions of the victim's employment and create an abusive working environment.'"[37] "Conditions" here has the same broad meaning that it does under disparate impact analysis.[38] For the plaintiff to prevail, he must establish that the harassing circumstances are "severe or pervasive enough to create an objectively hostile or abusive work environment—an environment that a reasonable person would find hostile or abusive."[39] Hostile work environment claims can be brought as disparate treatment claims, in which case the plaintiff needs to prove discriminatory intent, or as a type of disparate impact claim, which is more common in suits against English-only rules.[40]

---

31. Garcia v. Spun Steak Co., 998 F.2d 1480, 1484–85 (9th Cir. 1993) (quoting 42 U.S.C. § 2000e-2(a)(1)).

32. Meritor Sav. Bank v. Vinson, 477 U.S. 57, 64 (1986) (quoting L.A. Dep't of Water & Power v. Manhart, 435 U.S. 702, 707 n.13 (1978)).

33. Faragher v. City of Boca Raton, 524 U.S. 775, 786 (1998) (quoting Oncale v. Sundowner Offshore Servs., Inc., 523 U.S. 75, 78 (1998)).

34. *E.g.*, *Garcia*, 998 F.2d at 1485.

35. 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2006).

36. *Id.* § 2000e-2(k)(1)(A)(ii).

37. Harris v. Forklift Sys., Inc., 510 U.S. 17, 21 (1993) (quoting Meritor Sav. Bank, FSB v. Vinson, 477 U.S. 57, 65, 67 (1986)) (internal citations omitted).

38. *Cf. Garcia*, 998 F.2d at 1485–86 (applying the broad definition of "conditions" that the Supreme Court used in *Meritor Savings Bank* to a disparate impact claim predicated on burdensome conditions).

39. *Harris*, 510 U.S. at 21.

40. Maldonado v. City of Altus, 433 F.3d 1294, 1304 (10th Cir. 2006). Though it is not relevant to the purposes of this Note, it should be noted that the remedies for disparate impact claims are less extensive than those for disparate treatment claims. *Id.*

The EEOC issued guidelines about disparate impact and hostile work environment claims concerning English-only rules in 1980.[41] The guidelines differentiate between blanket English-only rules that apply at all times and those that better conform to the necessities of the job by having a more limited application. The EEOC presumes that a blanket rule is "a burdensome term and condition of employment," meaning that a plaintiff who sues an employer because of that rule will have automatically made out a prima facie case of disparate impact.[42] The guidelines also acknowledge that English-only rules that apply at all times may create a hostile work environment for employees of non-U.S. national origin.[43] The EEOC is less critical of rules that do not apply at all times, but still advises courts that employers should have to show that even these policies are justified by business necessity.[44] Courts have varied in the value they place on these EEOC guidelines.[45]

Beyond the disparate impact and hostile work environment claims endorsed by the EEOC guidelines, plaintiffs may be able to claim that English-only rules constitute overt systemic disparate treatment.[46] An overt systemic disparate treatment claim can be brought against a policy that facially discriminates against a protected class and has widespread effects.[47] If a plaintiff can show that an employer's policy is facially discriminatory, the only way for an employer to avoid liability is to establish that the discrimination relates to a bona fide occupational qualification (BFOQ).[48] The Supreme Court interprets the BFOQ defense very narrowly.[49] In order to carry its burden, an employer must show that the discriminated-against trait "relate[s] to the 'essence' . . . or to the 'central mission of the employer's business.'"[50] That task is significantly more difficult than establishing business necessity under the disparate impact framework.[51]

41. 29 C.F.R. § 1606.7 (2012).

42. *Id.* § 1606.7(a).

43. *Id.*

44. 29 C.F.R. § 1606.7(b).

45. *Compare* Garcia v. Spun Steak Co., 998 F.2d 1480, 1489 (9th Cir. 1993) (declining to follow the guidelines), *with* EEOC v. Synchro-Start Prods., Inc., 29 F. Supp. 2d 911, 914 (N.D. Ill. 1999) (showing significant deference to the guidelines).

46. I learned this apt term in Professor Joseph Fishkin's employment discrimination course and was surprised to find it had not been adopted elsewhere.

47. MICHAEL J. ZIMMER ET AL., CASES AND MATERIALS ON EMPLOYMENT DISCRIMINATION 115–16 (7th ed. 2008).

48. 42 U.S.C. § 2000e-2(e)(1) (2006); UAW v. Johnson Controls, Inc., 499 U.S. 187, 200 (1991).

49. *See Johnson Controls*, 499 U.S. at 201 (noting that "[t]he BFOQ defense is written narrowly, and this Court has read it narrowly").

50. *Id.* at 203 (internal citations omitted).

51. *Id.* at 198.

II. Stories of English-Only Rules and Accent Discrimination

Most lawsuits concerning English-only rules and accent discrimination are personal stories of struggles to feel accepted in the workplace. Below is a selection of a few of these stories, some of which courts found to be compelling enough to warrant relief, some of which they did not.

A. English-Only Rules

Jessie Kania was a Polish immigrant who worked as a housekeeper for a church.[52] She was bilingual and mostly spoke English on the job, but she would occasionally speak in Polish.[53] After Kania had been working at the church for five years, it enacted a rule making English the "official language" of the church and banning employees from speaking Polish during business hours.[54] Kania objected to the rule and maintained that the church "did not have the right to prevent her from speaking her native language at work," and she was fired a few weeks later.[55] In her lawsuit for national origin discrimination, the court did not address her allegation that "the English-only policy was a blanket rule that applied at all times during business hours, including when the Church's employees were at lunch, on break, and in non-public areas."[56] Instead, the court was content to credit the church's justification that "it is offensive and derisive to speak a language which others do not understand,"[57] and that an English-only rule was necessary "to improve interpersonal relations at the Church, and to prevent Polish-speaking employees from alienating other employees, and perhaps church members."[58]

Priscilla Garcia and Maricela Buitrago worked at Spun Steak, a poultry- and meat-product distributor that employed primarily bilingual Hispanic workers.[59] These employees generally spoke to each other in Spanish, until management received a complaint that they were harassing non-Spanish-speaking employees in Spanish.[60] Spun Steak then issued a rule that only English could be spoken while working (although Spanish could still be spoken during lunch and breaks), as well as a rule forbidding all offensive remarks.[61] Spun Steak explained that the English-only rule was enacted partly to "enhance worker safety because some employees who did not understand Spanish claimed that the use of Spanish distracted them while

52. Kania v. Archdiocese of Phila., 14 F. Supp. 2d 730, 731 (E.D. Pa. 1998).
53. *Id.*
54. *Id.*
55. *Id.* at 732.
56. *Id.* at 731.
57. *Id.*
58. *Id.* at 736.
59. Garcia v. Spun Steak Co., 998 F.2d 1480, 1483 (9th Cir. 1993).
60. *Id.*
61. *Id.*

they were operating machinery."[62] After Garcia and Buitrago received warnings for speaking Spanish at work, they filed a lawsuit alleging that the rule had a disparate impact against them on the basis of their national origin.[63] The court emphasized that the employees did not have a protected right to "express their cultural heritage at the workplace," and that they could "readily comply with the English-only rule."[64] It also disregarded the employees' allegation that they might involuntarily violate the English-only rule because of uncontrollable code-switching between Spanish and English.[65] Lastly, the court did not credit the employees' claim—supported by the EEOC guidelines—that such a rule inherently creates a hostile work environment for them because of their national origin.[66] The court instead demanded to see specific proof of this matter, which it found lacking in the plaintiffs' case.[67]

Tommy Sanchez was one of twenty-nine bilingual Spanish- and English-speaking employees employed by the city of Altus, Oklahoma.[68] When Sanchez heard that the city was going to pass an English-only rule, he raised concerns about the rule with the city's Street Commissioner.[69] The commissioner dismissed Sanchez's complaint and argued that Sanchez "would feel uncomfortable if another race would speak their native language in front of [him]."[70] Sanchez wrote a letter responding that:

> [W]e Hispanics are proud of our heritage and do not feel that our ability to communicate in a bilingual manner is a hindrance or an embarrassment. There has never been a time that because I spoke Spanish to another Spanish speaking individual, I was unable to perform our job duties and requirements.[71]

The city continued with its plans and passed a rule that "all work related and business communications during the work day shall be conducted in the English language."[72] The policy made two exceptions, one for "strictly private communications between co-workers" (but only during breaks, and only "if City property is not being used for the communication"), and the other for "strictly private communication between an employee and a family member" (but only if "the communications are limited in time and are not disruptive to the work environment").[73] The employees stated that the effect

62. *Id.*
63. *Id.* at 1483–85.
64. *Id.* at 1487.
65. *Id.* at 1488. For a discussion of code-switching, see *infra* subpart IV(B).
66. *Garcia*, 998 F.2d at 1488–89.
67. *Id.* at 1489.
68. Maldonado v. City of Altus, 433 F.3d 1294, 1298 (10th Cir. 2006).
69. *Id.* at 1299.
70. *Id.* (alteration in original) (internal quotation marks omitted).
71. *Id.*
72. *Id.* (emphasis omitted).
73. *Id.* (emphasis omitted).

of the rule was such that they could never speak Spanish around non-Spanish speakers, even during phone calls with family members.[74] In suing the city over its English-only policy, each plaintiff stated that the rule "reminds me every day that I am second-class and subject to rules for my employment that the Anglo employees are not subject to."[75] The plaintiffs also introduced evidence that they were consistently mocked by non-Hispanic employees because of this policy, and frequently reminded that they were forbidden from speaking Spanish.[76] The district court granted summary judgment against the plaintiffs' claims that the English-only rule caused a disparate impact and created a hostile work environment,[77] but the Tenth Circuit found that there were genuine issues of material fact and overturned the lower court's decision.[78]

Albert Estrada and Francisco Gracia were among a group of bilingual workers hired by Premier Operator Services to serve as telephone operators specifically because of their Spanish-speaking abilities.[79] However, to control the use of that Spanish, Premier posted a sign on the door of the building that read:

> *Absolutely* No Guns, Knives or Weapons of any kind are allowed on these Premises at any time! *English* is the official language of Premier Operator Services, Inc. All conversations on these premises are to be in English. Other languages may be spoken to customers who cannot speak English.[80]

The policy banned employees from speaking Spanish at all times, including during lunch and breaks.[81] For personal calls in Spanish, the employer installed a pay phone outside of the building.[82] The company president was also quoted as referring to his Hispanic employees as "wetbacks" multiple times.[83] Estrada and Gracia filed a complaint with the EEOC after Premier forced them to sign a memo stating that they agreed to the English-only policy.[84] Premier fired six employees who refused to sign the memo, and then fired Estrada and Gracia after they filed their complaint.[85] In a lawsuit brought by the EEOC on behalf of these employees, the court agreed that the

74. *Id.* at 1300.
75. *Id.* at 1301.
76. *Id.*
77. *Id.* at 1302.
78. *Id.* at 1316.
79. EEOC v. Premier Operator Servs., Inc., 113 F. Supp. 2d 1066, 1068 (N.D. Tex. 2000).
80. *Id.* at 1069. The court noted that it was "conspicuous[]" that the sign coupled the English-only policy with a prohibition on weapons. *Id.* at 1068–69.
81. *Id.* at 1069.
82. *Id.* at 1071.
83. *Id.*
84. *Id.* at 1069.
85. *Id.*

plaintiffs had shown that the employer enacted the rule with a discriminatory intent[86] and awarded damages to the employees.[87]

B. Accent Discrimination

Manuel Fragrante grew up in the Phillipines, where he learned English from an early age.[88] He served in the U.S. military during the Vietnam War. After the war, he continued his military training in Indiana and Kansas.[89] Fragante repeatedly earned "excellent" English language ratings from his superiors and never received complaints about his accent.[90] At the age of 60, he emigrated to Hawaii and, though he was old enough to retire, began looking for a job.[91] He applied to be a clerk at the City of Honolulu's Division of Motor Vehicles and Licensing, which required that he take an exam that tested "among other things, word usage, grammar and spelling."[92] Out of 721 test takers, Fragante's score was the highest.[93] But when he was interviewed for the position, his interviewers stated that they had a hard time understanding him because of his Filipino accent, and concluded that his accent "would interfere with his performance of certain aspects of the job."[94] At the trial for his disparate treatment claim against the DMV, two expert witnesses testified that even though Fragante had a heavy accent, his speech was comprehensible,[95] but that because of a history of discrimination against foreign accents like his, listeners may "turn off" and not understand him.[96] However, in affirming the trial judge's decision to dismiss Fragante's complaint, the court explained that there was nothing wrong with making an "*honest* assessment" of a candidate's "oral communication skills,"[97] and credited the district court's finding that Fragante "ha[d] a difficult manner of pronunciation,"[98] even though there was only one occasion during the trial when the judge had to ask Fragante to clarify a statement.[99]

Ramzy Salem was an immigrant from Palestine who had worked at La Salle High School for thirteen years, eleven of which as the chairperson of

86. *Id.* at 1071.

87. *Id.* at 1077–78.

88. Mari J. Matsuda, *Voices of America: Accent, Antidiscrimination Law, and a Jurisprudence for the Last Reconstruction*, 100 YALE L.J. 1329, 1334 (1991).

89. *Id.* at 1334–35.

90. *Id.* at 1335.

91. *Id.*; Fragante v. City of Honolulu, 888 F.2d 591, 593 (9th Cir. 1989).

92. *Fragante*, 888 F.2d at 593.

93. *Id.*

94. *Id.* at 593–94.

95. *Id.* at 595.

96. Matsuda, *supra* note 88, at 1337.

97. *Fragante*, 888 F.2d at 593, 596–98.

98. *Id.* at 598.

99. *See* Matsuda, *supra* note 88, at 1338 & n.28 (examining the transcript from the trial).

the school's mathematics department.[100] However, when the school's administration changed before what would have been his fourteenth year at the school, the new principal told Salem that, in addition to alleged insufficiencies in his teaching abilities, "[l]anguage difficulties . . . hinder[ed] [his] ability to function" to the point that the principal had decided not to offer him a new contract.[101] In its defense against Salem's lawsuit for disparate treatment on the basis of national origin, the school presented several reasons aside from Salem's accent that caused it not to offer him a new contract, but it also argued that Salem should have taken steps to lessen the effects of his accent.[102] The court credited these as legitimate, nondiscriminatory reasons for the school's actions, and found that Salem had not shown that he was discriminated against because of his Palestinian nationality, meaning that he could not establish that the school's proffered reasons were a pretext.[103]

Patricia Lee was born in China and received her medical degree from the National Taiwan University College of Medicine, then moved to the United States and practiced medicine at the Veterans Administration Medical Center in Pennsylvania.[104] She worked there as a physician for fifteen years, during which time she regularly requested, but was denied, promotion.[105] Her superiors frequently complained about her foreign accent.[106] One supervisor would get angry with her because he could not understand her, and another supervisor would not talk to her unless she was with someone who could act as an interpreter for her.[107] When Lee sued the hospital for race and national origin discrimination, the hospital claimed that she did not receive a promotion because her credentials from the Taiwanese medical school were not adequate.[108] The court determined that this explanation was a pretext for national origin discrimination, as degrees from the school Lee attended qualify its graduates to sit for licensing exams in the U.S.[109] The court also emphasized that Lee's accent—while "quite noticeable"—did not hinder her ability to communicate and should not have been considered in decisions about Lee's promotion.[110]

---

100. Salem v. La Salle High Sch., No. CV 82-0131-ER, 1983 U.S. Dist. LEXIS 18145, at *2 (C.D. Cal. Mar. 29, 1983).

101. *Id.* at *3.

102. *Id.* at *4–7.

103. *Id.* at *8–11.

104. Lee v. Walters, CIV. A. No. 85-5383, 1988 WL 105887, at *1 (E.D. Pa. Oct. 11, 1988).

105. *Id.*

106. *Id.* at *4.

107. *Id.*

108. *Id.* at *5.

109. *Id.* at *7.

110. *Id.*

Casserene Cassells, a woman of Jamaican ancestry, worked as a nurse at a hospital.[111] Cassells claimed she was subjected to frequent abuse at the hands of her supervisor. Along with several racially charged remarks, Casells' supervisor told her that she "did not like Jamaicans," and that "she had previously beaten a Jamaican woman and would do the same to [Cassells] if she did not follow orders."[112] The supervisor also ordered Cassells to "get rid of her Jamaican accent."[113] Cassells sued the hospital for disparate treatment on the basis of her race and national origin, and the court found that her claims were sufficient to survive the hospital's motion for summary judgment.[114]

III. The Shortcomings of the Courts' Decisions

The above cases expose a set of common mistakes that courts make in their reasoning in English-only rule and accent discrimination cases. First, it is quite likely that some of these courts simply do not place much value on foreign languages and accents. This is admittedly a somewhat inflammatory argument, but there is a good deal of evidence to support it. Regardless of whether they fall prey to devaluing foreign languages and accents or not, courts certainly underestimate the harm done to employees when they are denied jobs because of their accents or told they cannot speak their native languages any time they are working.

A. Devaluing Other Languages and Accents

It is quite possible that courts so often rule against plaintiffs in English-only rule and accent discrimination cases because—perhaps on a subconscious level—they do not place much value on foreign languages and accents. There is a long history of American hostility to foreign immigration.[115] In more recent years, this hostility has been particularly strong against Hispanics, which has led to an aversion to the Spanish

111. Cassells v. Univ. Hosp. at Stony Brook, 740 F. Supp. 143, 144 (E.D.N.Y. 1990).

112. *Id.* at 146.

113. *Id.*

114. *Id.* at 148.

115. *See* Paxton & Mughan, *supra* note 1, at 549–50 ("Even the United States, itself perhaps the archetypal immigrant society, has a long history of prejudice against newcomers to its shores.").

language.[116] Many English-only rules appear to have been the result of that sentiment.[117]

Through their questionable choice of language and lack of scrutiny of employers' stated business reasons for these policies, courts show that perhaps they too think that Spanish and other foreign languages are not worth protecting. Alfredo Mirandé observes that courts sometimes use problematic language in discussing English-only policies. He points out that in English-only rule cases, courts often refer to employees being "caught" or "overheard" speaking Spanish, and "voluntarily" speaking Spanish even though "they were 'capable' of speaking English."[118] This choice of language—as opposed to simply stating that "the employee spoke Spanish in violation of the policy"—indicates that courts may sometimes feel that there is something inherently wrong with speaking Spanish at work. Mirandé also argues that many English-only rules are effectively just "no Spanish" rules.[119] By crediting employers' flimsy business-necessity justifications for such rules, courts show that they do not place much value on employees' interest in being able to speak Spanish in the workplace.

Other business justifications accepted by courts expose their possible bias against foreign languages more generally. The court in *Kania v. Archdiocese of Philadelphia*[120] accepted the employer's statement that "it is offensive and derisive to speak a language which others do not understand."[121] Note that the employer here did not say anything about the content of that speech—just that the foreign language itself was offensive. In another case, an employer justified an English-only policy by stating that speaking Spanish was "very rude," and that refraining from speaking a language in front of someone who does not understand it was a matter of

116. *See, e.g.*, JAMES CRAWFORD, HOLD YOUR TONGUE: BILINGUALISM AND THE POLITICS OF "ENGLISH ONLY" 150–51 (1992) (discussing the hostility of the chairman of the U.S. English movement to Hispanics and the Spanish language); Alfredo Mirandé, "En la Tierra del Ciego, El Tuerto es Rey" *("In the Land of the Blind, the One Eyed Person is King"): Bilingualism as a Disability*, 26 N.M. L. REV. 75, 102–03 (1996) (observing that "speaking Spanish in the United States has been devalued historically" and illustrating that by pointing to the historical prevalence of "No Spanish" rules in schools throughout the Southwest and the punishment of "Spanish detention").

117. *See, e.g.*, Saucedo v. Bros. Well Serv., Inc., 464 F. Supp. 919, 921–22 (S.D. Tex. 1979) (holding against an employer who had a rule prohibiting any "Mesican talk" at work and who assaulted a Mexican employee when he tried to defend a coworker who was fired for speaking Spanish); Maldonado v. City of Altus, 433 F.3d 1294, 1301 (10th Cir. 2006) (recounting that the mayor of a town that instituted an English-only rule for municipal employees was quoted as calling Spanish "garbage"; the mayor later "claim[ed] that he used the word *garble* and was misquoted"); EEOC v. Premier Operator Servs., Inc., 113 F. Supp. 2d 1066, 1069 (N.D. Tex. 2000) (noting the stigmatizing effect of including an English-only rule directly following and on the same sign as a prohibition on weapons).

118. Mirandé, *supra* note 116, at 103.

119. *Id.* at 85–86.

120. 14 F. Supp. 2d 730 (E.D. Pa. 1998).

121. *Id.* at 731.

"common courtesy."[122] The trial court found no problem with that justification.[123] The court in *Garcia v. Spun Steak*[124] did not address the questionable reasoning behind the employer's justification that hearing Spanish was "distract[ing]."[125] That courts would not scrutinize employers' claims that speaking a foreign language is "offensive," "rude," and "distracting" says a great deal about their opinions of foreign languages.

Judicial treatment of accent discrimination cases also reveals a possible tendency to hold foreign accents in low esteem. For instance, even though the district court judge only had to ask Manuel Fragante to clarify one of his statements in the entire trial,[126] the judge stated that Fragante had a "difficult manner of pronunciation."[127] Furthermore, there is substantial evidence that listeners are vulnerable to being influenced by their biases in evaluating the speech of someone with a different manner of speaking than they have.[128] As one of Fragante's experts testified, people with biases against Filipinos would be likely to "turn off" when listening to someone like Fragante talk, and thus not be able to understand him.[129] There are marked differences in how people comprehend speech based on the perceived status of an accent. Those who speak with a low-status accent are forced by societal necessity to understand high-status accents, but those with high-status accents often cannot understand lower-status accents.[130] In this case, foreign accents hold this lower status, so many Anglo-Americans who speak with a so-called higher-status accent have trouble understanding foreign accents.[131] By not addressing these phenomena, there is a danger that judges are themselves falling victim to this sort of unconscious bias.

B. Underestimating the Harm

Even if courts are not biased against foreign languages and accents, their decisions show that many of them do not appreciate the harm that English-only rules and accent discrimination cause. In English-only cases,

---

122. Long v. First Union Corp. of Va., 86 F.3d 1151, 1996 WL 281954, at *1 & n.3 (4th Cir. 1996) (per curiam) (unpublished table decision).

123. Long v. First Union Corp. of Va., 894 F. Supp. 933, 941 (E.D. Va. 1995), *aff'd*, 86 F.3d 1151 (4th Cir. 1996).

124. 998 F.2d 1480 (9th Cir. 1993).

125. *Id.* at 1483.

126. Matsuda, *supra* note 88, at 1338 n.28.

127. Fragante v. City of Honolulu, 699 F. Supp. 1429, 1432 (D. Haw. 1987), *aff'd*, 888 F.2d 591 (9th Cir. 1989).

128. *See* Matsuda, *supra* note 88, at 1355, 1361 (analyzing the way in which those in power are perceived as speaking unaccented English, while low-status accents are perceived as foreign and unintelligible).

129. *Id.* at 1337.

130. *Id.* at 1352.

131. *See id.* (recounting an anecdote about a landlord who told the author that a foreign-accented neighbor could barely speak English, even though the author found that the neighbor was educated in the United States and could speak "perfect" English).

plaintiffs are almost always told that speaking English instead of Spanish is simply a "matter of choice," so the rules do them no harm.[132] In accent cases, courts generally find that limiting immigrants' employment opportunities because of their accents is not a serious harm, because they feel that immigrants can simply work to get rid of their accents.[133] Even when courts do acknowledge that accent-based hiring decisions or English-only rules harm the plaintiffs, they are still very deferential to employers' claims that these hiring decisions and language policies are justified by business necessity and thus rule against the plaintiffs.[134] Courts normally assume that these decisions were not made with any harmful discriminatory intent unless that intent is blatantly obvious from the facts of the case.[135]

In *Garcia v. Spun Steak*, the court found that Spun Steak's English-only rule did not have an adverse impact on the Spanish-speaking employees because "the rule is one that the affected employee can readily observe and nonobservance is a matter of individual preference."[136] The court also held that in order to show that the rule created a hostile work environment, the employees would have to make a specific factual showing of the hostile circumstances.[137] In doing so, the court dismissed as "conclusory" the plaintiffs' allegations that "the policy has contributed to an atmosphere of 'isolation, inferiority or intimidation,'" essentially ignoring the possibility that the existence of the English-only rule itself had a harmful impact.[138] The court in *Long v. First Union*[139] took a similar approach.[140] It framed the issue as a dispute over the "privilege" of speaking in one's "native tongue" at work, the denial of which did not amount to an adverse impact.[141] In *Kania v. Archdiocese of Philadelphia*, the court also emphasized that since the bilingual plaintiff "could have readily complied with the English-only rule,"

132. *See, e.g.*, Garcia v. Spun Steak Co., 998 F.2d 1480, 1487 (9th Cir. 1993) ("The bilingual employee can readily comply with the English-only rule and still enjoy the privilege of speaking on the job.").

133. *See, e.g.*, Salem v. La Salle High Sch., No. CV 82-0131-ER, 1983 U.S. Dist. LEXIS 18145, at *7 (C.D. Cal., Mar. 29, 1983) ("Plaintiff's language difficulties were not an immutable characteristic of his national origin and could have been improved if plaintiff had been willing to do so.").

134. *See infra* notes 151–53 and accompanying text.

135. *See, e.g.*, *supra* notes 104–10 and accompanying text.

136. *Garcia*, 998 F.2d at 1487 (quoting Garcia v. Gloor, 618 F.2d 264, 270 (5th Cir. 1980)) (internal quotation marks omitted).

137. *See id.* at 1489 (refusing to create a per se rule that English-only rules always create a hostile work environment and concluding that the employees had not raised sufficient evidence of a hostile atmosphere).

138. *Id.*

139. 894 F. Supp. 933 (E.D. Va. 1995), *aff'd*, 86 F.3d 1151 (4th Cir. 1996).

140. *Id.* at 941.

141. *Id.*

it did not cause her any harm.[142] All of these courts dismissed the idea of some right of employees to "express their cultural heritage" at work.[143]

There are similar problems in accent discrimination cases. For instance, in *Salem v. La Salle High School*,[144] the court did not recognize that being unnecessarily criticized for one's foreign accent could itself be harmful.[145] It focused its inquiry on finding concealed discriminatory intent behind the school's reasons for not renewing Salem's contract[146] instead of seeing that the school's negative opinion of the plaintiff's accent could itself be discriminatory if it was not supported by business necessity.

In addition to underestimating the harm caused by English-only rules and accent discrimination, courts tend to ignore the risk that these rules and decisions are being made with the intent to inflict harmful discrimination on employees because of their foreign origin. Granted, courts should not start from the assumption that employers have a discriminatory intent in enacting English-only rules or making employment decisions based on accents. But courts should also not limit their inquiries into the possibility of discriminatory intent only to cases of blatant discrimination like that in *EEOC v. Premier Operator Services*[147] and *Cassells v. University Hospital at Stony Brook*.[148] Prior to its decision in *Garcia v. Spun Steak*, the Ninth Circuit recognized the threat of national origin discrimination lurking behind English-only rules and accent preferences. In *Gutierrez v. Municipal Court*,[149] the court stated that "[b]ecause language and accents are identifying characteristics, rules which have a negative effect on bilinguals, individuals with accents, or non-English speakers, may be mere pretexts for intentional national origin discrimination."[150]

Courts have ignored this risk in deferring completely to employers' proffered business justifications for English-only rules and accent-based employment decisions. In *Garcia v. Spun Steak*, the court accepted Spun Steak's assertion that its English-only rule was put in place because Spanish-speaking employees had insulted other employees, and neglected to analyze the significance of a second rule that Spun Steak issued that prohibited all offensive remarks.[151] That rule alone could have resolved the employee

142. 14 F. Supp. 2d 730, 736 (E.D. Pa. 1998).

143. *Id.*; *Garcia*, 998 F.2d at 1487; *Long*, 894 F. Supp. at 941.

144. No. CV 82-0131-ER, 1983 U.S. Dist. LEXIS 18145 (C.D. Cal., Mar. 29, 1983).

145. *See id.* at *10–11 (finding that the plaintiff had failed to establish national origin discrimination in not having his contract renewed, but failing to recognize any ties between national origin and accent).

146. *Id.* at *3–4.

147. 113 F. Supp. 2d 1066 (N.D. Tex. 2000).

148. 740 F. Supp. 143 (E.D.N.Y. 1990); *see supra* notes 79–87, 111–14 and accompanying text.

149. 838 F.2d 1031 (9th Cir. 1988), *vacated as moot* 490 U.S. 1016 (1989).

150. *Id.* at 1039 (citing Tom McArthur, Comment, *Worried About Something Else*, 60 INT'L J. SOC. LANGUAGE 87, 90–91 (1986)).

151. Garcia v. Spun Steak Co., 998 F.2d 1480, 1483 (9th Cir. 1993).

conflicts, and it applied equally to everyone at the company, thus avoiding the harm caused by the English-only rule singling out Spanish-speaking employees.[152] The *Kania* court was similarly lenient in assessing the validity of the employer's business necessity justification that the English-only rule was needed "to improve interpersonal relations at the Church, and to prevent Polish-speaking employees from alienating other employees, and perhaps church members themselves."[153] By not making the slightest inquiry into why the church felt that Kania's Polish was hurting interpersonal relations or "alienating other employees," the court ignored the possibility that those justifications were pretexts for harmful discrimination.

IV. The Harm of English-Only Rules and Accent Preferences to National Origin Groups

In order to correct the mistakes discussed in the previous Part, courts must understand the damage that English-only rules and accent discrimination can inflict upon workers of foreign national origin. Explaining that harm requires illustrating the link between an employee's national origin and his language and accent, and the importance of that language and accent to the employee's identity. To understand that harm, courts must also reexamine their treatment of foreign languages and accents as mutable. Once a court appreciates the significance that accent and language can have to a foreign employee's sense of self, it is easy to see how detrimental it can be when employers force employees to cover those aspects of their identities.

A. The Significance of Language and Accent to Identity

1. The Connection Between National Origin, Language and Accent.— In the first major decision concerning an English-only rule, *Garcia v. Gloor*, [154] the court found that "[n]either [Title VII] nor common understanding equates national origin with the language that one chooses to speak."[155] Perhaps common understanding has changed, but it is now hard to dispute that language, as well as accent, is directly connected to national origin.

Since *Garcia v. Gloor*, courts have recognized the link between national origin and language. The *Gutierrez* court acknowledged that "language is an important aspect of national origin."[156] The court went on to state that "[t]he

152. *Id.* at 1483; *see* Answering Brief at 32–34, *Garcia*, 998 F.2d 1480 (9th Cir. 1993) (No. 91-16733).

153. Kania v. Archdiocese of Phila., 14 F. Supp. 2d 730, 736 (E.D. Pa. 1998).

154. 618 F.2d 264 (5th Cir. 1980).

155. *Id.* at 268.

156. Gutierrez v. Mun. Court, 838 F.2d 1031, 1039 (9th Cir. 1988), *vacated sub nom. as moot* 490 U.S. 1016 (1989).

mere fact that an employee is bilingual does not eliminate the relationship between his primary language and the culture that is derived from his national origin."[157] The Ninth Circuit later reiterated this sentiment, finding that "language is a close and meaningful proxy for national origin."[158] Courts have also accepted that accent is tied to national origin. In *Fragante v. City & County of Honolulu*, the court found that "[a]ccent and national origin are obviously inextricably intertwined in many cases."[159]

Commentators have provided additional support for these connections. As Professor Perea put it, "[p]rimary language, like accent, is closely correlated and inextricably linked with national origin."[160] Janet Ainsworth summed up this retort to *Garcia v. Gloor* nicely, writing that "*it is beyond dispute* that, for many individuals, their mother tongue is a function of their ethnic background."[161]

2. The Importance of Language and Accent to One's Sense of Self.— Furthermore, a person's language and accent have a close connection to his national origin identity, or his sense of self that derives from his national origin. Therefore, any limitations placed on a person because of his accent or native language harm his national origin identity, placing an impermissible burden on him because of his national origin.

Both courts and commentators have recognized the importance of language to identity. In *Gutierrez v. Municipal Court*, the Ninth Circuit found that "[t]he cultural identity of certain minority groups is tied to the use of their primary tongue."[162] In his dissent from the denial to rehear *Garcia v. Spun Steak* en banc, Judge Reinhardt recognized that an immigrant's "native language remains an important manifestation of his ethnic identity and a means of affirming links to his original culture."[163] Commentators have observed that language "touches the sense of belonging, and undoubtedly that sense is vital to every person's identity and self-esteem,"[164] and that

---

157. *Id.* (citing Kenneth L. Karst, *Paths to Belonging: The Constitution and Cultural Identity*, 64 N.C. L. REV. 303, 351–57 (1986)).

158. Yniguez v. Arizonans for Official English, 69 F.3d 920, 947–48 (9th Cir. 1995) (en banc), *vacated sub nom.* Arizonans for Official English v. Arizona, 520 U.S. 43 (1997).

159. Fragante v. City of Honolulu, 888 F.2d 591, 596 (9th Cir. 1989). The Ninth Circuit reaffirmed its recognition of this connection fifteen years later. *See* Fonseca v. Sysco Food Servs. of Az., Inc., 374 F.3d 840, 849 n.4 (9th Cir. 2004)

160. Juan F. Perea, *English-Only Rules and the Right to Speak One's Primary Language in the Workplace*, 23 U. MICH. J. L. REFORM 265, 276 (1990).

161. Janet Ainsworth, *Language, Power, and Identity in the Workplace: Enforcement of 'English-Only' Rules by Employers*, 9 SEATTLE J. FOR SOC. JUST. 233, 237 (2010) (emphasis added).

162. Gutierrez v. Mun. Court, 838 F.2d 1031, 1039 (9th Cir. 1988), *vacated as moot* 490 U.S. 1016 (1989).

163. Garcia v. Spun Steak Co., 13 F.3d 296, 298 (9th Cir. 1993) (Reinhardt, J., dissenting).

164. *See* Karst, *supra* note 157, at 356 (discussing language in the context of bilingual education).

"deprivations in relation to language deeply affect identity."[165] The close tie between language and identity has been studied by anthropologists, sociologists, and sociolinguists.[166]

There is a particularly strong scholarship linking the Spanish language to Latino identity. Spanish has been deemed "an intractable part of the Latino culture, representing one of the ties of Spanish-speaking persons to their ancestors' or their own place of origin."[167] In exploring the importance of Spanish to Latino professionals, Maria Chávez observes that embracing one's national origin identity, to which language is essential, is "critical to survival."[168] Through survey work, she found that over a third of Latino lawyers still speak Spanish "on social occasions," and feel that the language "is very important to their identity."[169] Alfredo Mirandé discusses his personal experiences with the bond that Spanish creates between Mexican-Americans, which "transcend[s] educational and class differences."[170] The Supreme Court even weighed in on the subject, observing that the Spanish language is used by many Latinos "to define the self."[171]

Accent is also tied to our sense of self. Mari Matsuda describes how our accents carry the stories of who we are,[172] and asserts that "[t]he way in which we speak reflects self, personhood, identity."[173] She relates the story of how during a discussion about accent discrimination cases, a student's comment that "I don't see how they can come to our place and tell us we can't talk the way we talk" brought her to tears.[174] It made her recognize that

165. Myres S. McDougal et al., *Freedom from Discrimination in Choice of Language and International Human Rights*, 1 S. ILL. U. L.J. 151, 151 (1976).

166. *See* Ainsworth, *supra* note 161, at 245 n.50 (listing sources from various disciplines that have explored the subject).

167. Christian A. Garza, Case Note, *Measuring Language Rights Along a Spectrum*, 110 YALE L.J. 379, 382 (2000). Garza goes on to point out that "[t]his experience is not limited to Latinos; the connection is equally strong among other language minority groups." *Id.*

168. MARIA CHÁVEZ, EVERYDAY INJUSTICE: LATINO PROFESSIONALS AND RACISM 39 (2011).

169. *Id.* at 41. This number is more significant when one realizes that most of that one third probably belonged to the 50% of survey respondents who spoke English as a second language. *See id.* at 50 ("Almost half of the Latino attorneys in this study spoke English as a second language. Close to 40 percent of these Latino lawyers still speak Spanish, and language is a key link to culture and community.").

170. Mirandé, *supra* note 116, at 92 n.147.

171. Hernandez v. New York, 500 U.S. 352, 363–64 (1991) (examining whether the use of language to strike jurors could be considered race-based discrimination); *see also* Christopher David Ruiz Cameron, *How the García Cousins Lost Their Accents: Understanding the Language of Title VII Decisions Approving English-Only Rules as the Product of Racial Dualism, Latino Invisibility, and Legal Indeterminacy*, 85 CALIF. L. REV. 1347, 1353–54 (1997) ("[T]he Spanish language is central to Latino identity.").

172. Matsuda, *supra* note 88, at 1329.

173. *Id.* at 1388.

174. *Id.* at 1391 (internal quotation marks omitted). It is not clear from Matsuda's article, but I assume that the student was referring to the case *Kahakua v. Friday*, 876 F.2d. 896, 1989 WL61762, at *5 (9th Cir. 1989) (unpublished table decision), in which a court found that a news station had not discriminated when it chose not to hire the best qualified meteorologist for a weather forecaster

our accents reside in "the sacred places of the self."[175] Others have observed that when someone learns a new language, he must "giv[e] up part of [his] culture," of which "the last vestige" may be his accent.[176]

Recognizing the value of language and accent to one's sense of self, and their connection to national origin, establishes that actions that harm someone because of his foreign language or accent should be considered national origin discrimination under Title VII. The way in which English-only rules and accent preferences harm immigrant employees on the basis of their accents and native languages will be further explained in subpart IV(C), but first it is necessary to address the problematic issue of the immutability of these traits.

B. Immutability

There is considerable debate about whether Title VII should only protect immutable traits.[177] I do not intend to enter that debate here. Instead, I would like to briefly point out that even if Title VII should only protect immutable traits, language and accents are not as mutable as courts frequently assume. For instance, in *Garcia v. Spun Steak*, the court stated that "[i]t is axiomatic that 'the language a person who is multi-lingual elects to speak at a particular time is . . . a matter of choice.'"[178] This is a slightly absurd statement for a Ninth Circuit court to make, given that only a few years earlier, the Ninth Circuit in *Gutierrez* had recognized the inextricable link between language and a person's national origin identity and found that complying with an English-only rule was not "a matter of personal preference."[179] However, taking it at its word, there is a great deal of evidence that shows that the court's opinion about language, far from being "axiomatic," is likely not even true. The previous supbart described the integral—and perhaps immutable—connection between language, accent, and identity. This subpart will address the more scientific links that a person has to his accent and native language that make them essentially immutable.

---

opening because it felt that his Hawaiian Creole accent would hinder him from performing the job well.

175. Matsuda, *supra* note 88, at 1391.

176. Thomas J. Coates & Patricia M. Regdon, *Thrice: A Technique for Improving the American English Language Delivery of Non-Native Speakers*, 8 TESOL Q. 363, 369 (1974).

177. *Compare* Sharona Hoffman, *The Importance of Immutability in Employment Discrimination Law*, 52 WM. & MARY L. REV. 1483, 1488 (2011) (arguing for immutability but acknowledging its flaws), *with* Perea, *supra* note 22, at 866–67 (stating that the "presence or absence of mutability should not be relevant in fundamental matters of individual identity, such as ethnicity"), *and* Peter Brandon Bayer, *Mutable Characteristics and the Definition of Discrimination Under Title VII*, 20 U.C. DAVIS L. REV. 769, 839 (1987) (taking issue with the presumption that the importance of a trait is dependent upon how easily someone can change it).

178. Garcia v. Spun Steak Co., 998 F.2d 1480, 1487 (9th Cir. 1993) (quoting Garcia v. Gloor, 618 F.2d 264, 270 (5th Cir. 1980)).

179. Gutierrez v. Mun. Ct., 838 F.2d 1031, 1039–41 (9th Cir. 1988), *vacated as moot* 490 U.S. 1016 (1989).

There are various psychological and psycholinguistic ties between a person and his native language that make his use of that language effectively immutable. When a person learns a language when he is young, "it forms an immutable perspective and understanding," of which he likely cannot "consciously purge [himself]."[180] A person cannot change the "neurological processes" controlling that language "that [have] been set in place from a very early age."[181] The most significant result of these neurological phenomena is code-switching. Code-switching refers to the involuntary use of one's native language when speaking English.[182] It can happen to bilingual speakers quite frequently,[183] even if they have a negative attitude toward inadvertently shifting into their native language.[184] The court in *EEOC v. Premier Operator Services* recognized the legitimacy of the science behind code-switching and endorsed the testimony of an expert who stated that because of code-switching, the use of Spanish could not simply be "turned off."[185]

Scientific factors also affect the mutability of foreign accents. Mari Matsuda states that in issuing a guideline about accent discrimination, the EEOC relied on "evidence that it is nearly impossible for an adult to eliminate their natural accent."[186] She cites linguistic studies showing that when people learn second languages after childhood, they can almost never learn to speak those languages without an accent,[187] and that when non-native English speakers do try to speak without an accent, they often overcorrect and speak in a way that would be unnatural to a native speaker.[188] Matsuda also examines some rigorous techniques for teaching non-native English speakers to speak without an accent, which she finds "daunting and degrading."[189]

---

180. BILL PIATT, LANGUAGE ON THE JOB: BALANCING BUSINESS NEEDS AND EMPLOYEE RIGHTS 121 (1993).

181. *Id.*

182. EEOC v. Premier Operator Servs., Inc., 113 F. Supp. 2d 1066, 1069–70 (N.D. Tex. 2000).

183. Nanda Poulisse & Theo Bongaerts, *First Language Use in Second Language Production*, 15 APPLIED LINGUISTICS 36 (1994).

184. *See* William C. Ritchie & Tej K. Bhatia, *Social and Psychological Factors in Language Mixing*, *in* THE HANDBOOK OF BILINGUALISM 336, 350 (Tej K. Bhatia & William C. Ritchie eds., 2004) (noting that bilinguals often apologize for code-mixed speech and promise improvement); Itesh Sachdev & Howard Giles, *Bilingual Accommodation*, *in* THE HANDBOOK OF BILINGUALISM 353, 359–60 (relating that speakers who find themselves unconsciously code-switching consider the practice to be "an unneeded and disturbing mixture of languages").

185. *Premier Operator Servs., Inc.*, 113 F. Supp. 2d at 1069–70.

186. Matsuda, *supra* note 88, at 1348–49.

187. *Id.* at 1349 n.74 (citing Michael H. Long, *Maturational Constraints on Language Development*, 12 STUD. SECOND LANGUAGE ACQUISITION 251 (1990)).

188. *Id.* at 1349 & n.73 (citing William Labov, *Excerpt from the Study of Language in its Social Context*, *in* SOCIOLINGUISTS: SELECTED READINGS 191–93 (J.B. Pride & Janet Holmes eds., 1972)).

189. *Id.* at 1349 n.74 (citing Coates & Regdon, *supra* note 176, at 363).

Thus, even if a court insists that Title VII only offers protection to immutable traits, a non-native English speaker can make a strong argument that his accent and the use of his native language are immutable. Even if a court disregarded the above findings, it should still be sympathetic to the serious harm caused to employees who are forced to cover these central aspects of their national origin identities.

C. *Enforced Covering of National Origin Identity*

In his seminal article, *Covering*, Kenji Yoshino introduced a new framework for evaluating the harm caused by discrimination.[190] Yoshino linked the effect that forced covering, or "coerced assimilation," has on gays and lesbians to similar harms inflicted upon women and racial minorities. In Yoshino's terminology, to cover is to hide or mute a quality of one's personality.[191] In the gay context, covering refers to the fact that "it is now permissible both to be gay and to say that one is gay, as long as one does not flaunt one's homosexuality."[192] The pressure to cover can force a person into "explicitly making a compromise about" an element of his or her identity that is tied to a protected trait.[193] In this way, covering can be a "severe burden."[194] Yoshino illustrates the "seriousness of the harm the covering demand inflicts" by pointing out that in the gay context, "certain acts denominated as covering, such as abstention from same-sex sodomy, might be constitutive of gay identity."[195] He also lists "muting linguistic difference" as race-based and "muting a pregnancy" as sex-based examples of covering that are "constitutive of identity"[196] Yoshino warns that "the contemporary forms of discrimination to which racial minorities and women are most vulnerable often take the guise of enforced covering."[197]

Yoshino mentions that "lapsing into Spanish" when an employer has an English-only rule and "speaking with an accent" are examples of failing to cover, but he does not explore those examples in detail, and he analyzes them as race-based covering in his article and as ethnicity-based covering in his book.[198] He does not discuss the idea of forced covering of national origin

190. Yoshino, *supra* note 13, at 781.
191. *Id.* at 772.
192. *Id.* at 838.
193. *Id.*
194. *Id.* at 837.
195. *Id.* at 781.
196. *Id.*
197. *Id.*
198. *Id.*; *see also* KENJI YOSHINO, COVERING: THE HIDDEN ASSAULT ON OUR CIVIL RIGHTS 137–39 (2006) (describing language as "an important aspect of ethnic identity" and asserting that "English-only statutes punish individuals not for knowing too little, but for knowing too much"). So far, other authors have only very briefly discussed English-only rules through the framework of covering. *See* Cristina M. Rodríguez, *Language Diversity in the Workplace*, 100 NW. U. L. REV. 1689, 1727 (2006) (referring to Yoshino's claim that enforced covering affects racial minorities and women when addressing the "assimilationist expectations" inherent in English-only cases);

characteristics. I argue that covering analysis provides an effective framework for appreciating the harms imposed by English-only rules and accent preferences upon the national origin identities of immigrant workers. These rules and decisions force people to cover certain traits that can be "constitutive" of their national origin identities.

The English-only rule and accent discrimination cases and associated scholarship illustrate that these rules and decisions are, in effect, demands to cover national origin traits. When Cassarene Cassells's supervisor told her to get rid of her Jamaican accent[199] and Patricia Lee's superiors scolded her about her Chinese accent,[200] they were ordering them to cover an aspect of their national origin identities. All English-only rules force employees who speak foreign languages to cover a major element of who they are that is tied to their national origins. Rules that apply at all times, like those in *EEOC v. Premier Operator Services* and *Maldonado v. City of Altus*, force the most severe covering, since employees are made to feel that a part of their identities is so devalued that it must be hidden at all times.

The sources also show the consequences of these commands to cover. The testimony in *Maldonado v. City of Altus* that the city's English-only rule "reminds me every day that I am second-class and subject to rules for my employment that the Anglo employees are not subject to"[201] illustrates the harm felt by employees from having to cover their national origin identities. The reaction of one Hispanic man to legislation that made English the official language of California is illustrative of the impact of forced covering: "You don't feel as free when you perceive this language limitation. This is the language in which we express ourselves. You have to hold part of you back. You feel less free than the rest of the people in this society."[202] English-only rules can be perceived as telling Hispanics that "to be included into the structures of this society they have to relinquish a part of their culture."[203] The effect of forced covering of foreign accents is equally harmful. Being forced to cover can make immigrants feel like they are "somehow unworthy because of the way [they] talk."[204] Forced covering of accents can also have the extreme effect of making those with foreign accents feel that they should not speak at all: "To tell people they cannot

---

L. Darnell Weeden, *The Less than Fair Employment Practice of an English-Only Rule in the Workplace*, 7 NEV. L.J. 947, 947–48 (2007) (mentioning Yoshino's ideas about assimilation as applied to English-only rules).

199. Cassells v. Univ. Hosp. at Stony Brook, 740 F. Supp. 143, 146 (E.D.N.Y. 1990).

200. Lee v. Walters, CIV. A. 85-5383, 1988 WL 105887, at *4 (E.D. Pa. Oct. 11, 1988).

201. Maldonado v. City of Altus, 433 F.3d 1294, 1301 (10th Cir. 2006).

202. Wendy Olson, *The Shame of Spanish: Cultural Bias in English First Legislation*, 11 CHICANO-LATINO L. REV. 1, 25 (1991).

203. *Id.*

204. Matsuda, *supra* note 88, at 1391.

express themselves in the way that comes naturally to them is to tell them they cannot speak."[205]

Granted, all members of modern societies are forced to cover certain personality traits that they consider to be components of who they are. The difference here is the depth of the connection between these aspects of identity and one's national origin, and the increased harm that comes from having to hide a part of that identity. Congress made a decision to protect people from employment discrimination on the basis of their national origin, and that protection should extend to the kind of personal harm that is done when employers force employees to cover their natural language or accents. Statements that Title VII does not grant an employee a "right to speak his or her native tongue while on the job"[206] miss the point entirely. The issue is not one of protecting a right to "express [one's] cultural heritage" on the job,[207] but of prohibiting employers from inflicting harm upon people by devaluing a protected trait—their national origin. English-only rules and accent discrimination force immigrant employees to have, in the language of *Brown v. Board of Education*,[208] "a feeling of inferiority as to their status in the community" by telling them they need to hide their national origin.[209] Thus, the forced covering caused by English-only rules and accent preferences results in significant burdens being placed upon immigrants purely on the basis of their national origins, meaning that those policies and decisions should be considered forbidden national origin discrimination under Title VII unless employers have a *legitimate* business justification.

Some commentators, most notably Richard Thompson Ford, may object to this reasoning on the grounds that it reduces a group to certain essential qualities, and then assigns those qualities to all members of the group.[210] Ford illustrates the pitfalls of this approach with an example of a newspaper op-ed that called Anita Hill "disingenuous" for complaining about Clarence Thomas's alleged sexual harassment.[211] The columnist wrote that because Hill was from a black, working-class, Southern background, she "perfectly understood" the context of Thomas's conduct and that it was not meant to harass her.[212]

While Ford does raise an important concern, connecting foreign language and accents to national origin is distinguishable from the

---

205. *Id.* at 1388.

206. *See, e.g.*, Long v. First Union Corp. of Va., 894 F. Supp. 933, 941 (E.D. Va. 1995), *aff'd*, 86 F.3d 1151 (4th Cir. 1996).

207. *Id.*

208. 347 U.S. 483 (1954).

209. *Id.* at 494.

210. *See* RICHARD THOMPSON FORD, RACIAL CULTURE: A CRITIQUE 74 (2005) (criticizing the potential for group-recognition claims to "decide for all members of the group what is to be deemed fundamental to the identity of the group").

211. *Id.*

212. *Id.*

problematic essentialization he describes. First-generation immigrants do, essentially by definition, speak a foreign language, and unless they had access to a particularly exceptional education, they speak English with a foreign accent. In that way, language and accent are inextricably linked with national origin, reducing the danger that those qualities will be unfairly assigned to members of the group. Furthermore, those immigrants, and more likely descendants of immigrants, who do not speak a foreign language or speak English with an accent are free to not place value on a native language or accent. The goal is not to essentialize and prescribe identity for all members of national origin minorities, but to recognize that for a significant number of them, language and accent are in fact an integral part of their identities because of their national origins, and forcing them not to express that identity does very real harm.[213] Those members of foreign national origin groups who do not consider language or accent to be an important part of their identities should not feel that banning discrimination of those traits is a prescription for what traits should be important to them or a statement that they are covering if they do not embrace those traits, but instead see that ban as simply a protection for the many members of their group who do feel a connection to those traits. Yoshino recognized this response to Ford's concerns when he wrote that:

> [W]e must not assume that individuals behaving in 'mainstream' ways are necessarily covering. My ultimate commitment is to autonomy as a means of achieving authenticity, rather than to a fixed conception of what authenticity might be.... [T]he demand[s] to conform to the mainstream . . . are the demands that most threaten our authenticity.[214]

Thus, what matters most is that members of these groups be protected from being forced to cover these aspects of their national origin identities, not that they all necessarily embrace those aspects.

V. The Necessary Changes to the Adjudication of Accent Discrimination and English-Only Rule Lawsuits

A better understanding of the harm that employers inflict with English-only rules and accent preferences should change the way that courts evaluate Title VII claims brought against these policies and decisions. In lawsuits against English-only rules, courts should accept the existence of the English-only rule as proving a prima facie case for disparate impact and, in many cases, for hostile work environment and overt systemic disparate impact claims. Courts should also scrutinize employers' business necessity justifications more closely. In accent-discrimination cases, courts should

---

213. A comparison could be drawn to sexual harassment law—just because some women are not offended by sexually harassing conduct does not mean that it should not be protected, and the harm that sexual harassment causes outweighs the danger that some women who are not offended by sexually harassing conduct might feel they are being told that they should find it offensive.

214. YOSHINO, *supra* note 198, at 190–91.

more carefully evaluate employers' proffered reasons for considering an employee's accent in making an employment decision.

A. Changes to the Treatment of Claims Against English-Only Rules

1. Disparate Impact Claims.—If courts still find that English-only rules are facially neutral, then plaintiffs will be restricted to bringing disparate impact claims. The EEOC guidelines state that English-only rules that apply at all times trigger an automatic presumption of disparate impact.[215] While the impact of blanket rules is more severe than that of rules that do not apply during breaks—of which the EEOC guidelines are not as critical—an appreciation of the inherent harm caused by English-only rules should expand the EEOC's presumption about blanket rules to also apply to rules that do not extend to breaks.[216] *Meritor Savings Bank, FSB v. Vinson*[217] established that terms and conditions of employment are to be interpreted broadly.[218] A rule that forces an employee to hide an important aspect of his national origin identity, even if it does not apply at all times, has a significant enough impact to affect that employee's terms and conditions of employment. Thus, when an employee challenges an English-only rule, the burden should automatically fall to the employer to give a compelling business justification for the rule.

Courts should also be stricter in their evaluation of employers' business justifications for the rules. They should not allow employers to force employees to hide their connection to their native languages without scrutinizing questionable justifications, such as other workers being distracted by hearing Spanish,[219] respect for customers to whom the employee is not even speaking,[220] the assertion that speaking a language that bystanders do not understand is "offensive and derisive,"[221] improving the

---

215. 29 C.F.R. § 1606.7(a) (2012).

216. *Id.* § 1606.7(b).

217. 477 U.S. 57 (1986).

218. *Id.* at 64; *see supra* Part I.

219. *See* Garcia v. Spun Steak Co., 998 F.2d 1480, 1483 (9th Cir. 1993) (failing to scrutinize an employer's statement that an English-only policy "would enhance worker safety because some employees who did not understand Spanish claimed that the use of Spanish distracted them while they were operating machinery").

220. *See* EEOC v. Sephora USA, LLC, 419 F. Supp. 2d 408, 417 (S.D.N.Y. 2005) (holding, without citing any evidence, that "[w]hen salespeople speak in a language customers do not understand, the effects on helpfulness, politeness and approachability are real and are not a matter of abstract preference").

221. *See* Kania v. Archdiocese of Phila., 14 F. Supp. 2d 730, 731, 736 (E.D. Pa. 1998) (accepting the defendant's justification that an English-only rule would "prevent Polish-speaking employees from alienating other employees, and perhaps church members themselves" without considering whether that idea was itself discriminatory).

English skills of employees,[222] and the general discomfort of employees at hearing coworkers speak Spanish.[223] Courts should also inquire more deeply into whether there was a nondiscriminatory alternative to the English-only rule, as there was with the "no offensive remarks" policy in *Garcia v. Spun Steak*.[224]

Despite the harm that English-only rules cause, employers should be able to show that they are justified by business necessity in certain circumstances. For instance, in one EEOC decision,[225] the EEOC determined that an oil refinery employer was justified in having an English-only rule that only applied in processing and laboratory areas and during emergencies. That policy is narrowly tailored to compelling workplace safety needs, and should survive the close scrutiny that courts should apply when taking covering harm into account. It is possible that the only claims that will be justified by business necessity are those that can be proved to be necessary for communication-based (as opposed to anti-distraction)[226] safety. That justification may be the only truly neutral reason for having an English-only rule, as it does not needlessly force minorities to conform to the native-English-speaking majority in the way that justifications such as "promot[ing] racial harmony"[227] and not "alienating"[228] others do.[229]

2. Hostile Work Environment Claims.—Recognizing the greater harm that English-only rules cause non-native English speakers should make courts more likely to credit plaintiffs' assertions that the rules create hostile work environments. The EEOC guidelines establish that these should be viable claims against English-only rules.[230] A better understanding of the harm caused by enforced covering reinforces the EEOC's guidelines. It should also lead courts to accept more than just the claims against blanket rules endorsed by the EEOC, since this perception clarifies that there is still

222. *See* Garcia v. Gloor, 618 F.2d 264, 267, 270 (5th Cir. 1980) (affirming, without scrutiny, the district court's holding that the employer's proffered justification that requiring only English would help employees improve their English constituted a valid justification of the rule).

223. *See* Barber v. Lovelace Sandia Health Sys., 409 F. Supp. 2d 1313, 1337–38 (D.N.M. 2005) (accepting employee discomfort at other employees speaking Spanish as a legitimate nondiscriminatory justification for an English-only policy without analyzing whether that actually constituted a discriminatory reason).

224. 998 F.2d at 1483. Granted, an employer may be able to show that a no-offensive-remark policy would not be an effective alternative if no supervisors speak the same foreign language as their employees and thus cannot monitor the employees' speech.

225. EEOC Decision No. 83-7, 31 Fair Empl. Prac. Cas. (BNA) 1861 (1983).

226. *Garcia*, 998 F.2d at 1483.

227. *Id.*

228. *See* Kania v. Archdiocese of Phila., 14 F. Supp. 2d 730, 736 (E.D. Pa. 1998).

229. There could be rare circumstances where racial harmony could justify an English-only rule, as in a workplace hostilely divided into two language minorities, where speaking English is actually a neutral ground for those groups and not just a way to force minorities to conform their identities to the majority.

230. 29 C.F.R. § 1606.7(a) (2012); *see supra* note 43 and accompanying text.

serious harm even when a rule only applies during work hours. Courts should recognize, as the court in *Maldonado v. City of Altus* did, that "the very fact that [an employer] would forbid Hispanics from using their preferred language could reasonably be construed as an expression of hostility to Hispanics" that can create a hostile work environment, especially if the employer cannot offer a good business justification for why the rule was necessary.[231] That understanding would change the result in cases like *Kania v. Archdiocese of Philadelphia*, where the court demanded that the plaintiff present specific evidence of how the rule created a hostile environment.[232] If the court appreciated the harm caused by the existence of the rule, it should have scrutinized whether the employer had a good reason for enacting the rule, and if not, inferred that its enactment expressed hostility towards the plaintiff and created a hostile work environment.

3. Overt Systemic Disparate Treatment Claims.—Lastly, if courts appreciate the full extent of the harm caused by English-only rules, plaintiffs should be able to attack the rules for causing overt systemic disparate treatment. This charge would go beyond even what the EEOC guidelines advise about claims against English-only rules. For a plaintiff to successfully bring an overt systemic disparate treatment claim, he would have to show that the employer's policy was discriminatory on its face. That the policy applies only to certain groups based on national origin supports a claim that the rule is facially discriminatory.[233] However, the plaintiff would also have to establish that the rule was enacted in a workplace where the only employees affected are national origin minorities; otherwise, it could not be said that the rule explicitly burdens this protected group. Showing that the policy was a blanket rule that was not tailored to the needs of the business would support finding that the rule discriminated against foreign national origin groups on its face. If plaintiffs could bring these claims, employers would have a heightened burden of proving business necessity, in that they would have to show that speaking only English on the job was a BFOQ. That would effectively mean that employers would almost always lose these cases, as an employer could rarely show that speaking only English was essential to the essence of a business.[234]

---

231. Maldonado v. City of Altus, 433 F.3d 1294, 1305 (10th Cir. 2006).

232. *Kania*, 14 F. Supp. 2d at 735–36.

233. *See* Perea, *supra* note 160, at 293 (equating "[d]iscrimination on the basis of primary language" with "discrimination on the basis of national origin . . . because of the very close correlation between primary language and national origin and the exclusive adverse impact of restrictions upon the use of primary languages other than English" and arguing that "[t]he intent necessary to show disparate treatment can be inferred from the existence of such exclusive adverse effects").

234. For example, a BFOQ could exist for an English-only rule that prohibited speaking foreign languages on the air in an English-language broadcast, even if it only affected certain employees of foreign national origin.

EEOC v. Premier Operator Services provides a good example of a case where this sort of claim would be available, since the only employees affected were Hispanic, and the employer's policy was written in a way that evinced a discriminatory intent. The outcome of the case would not change under the new framework, since in the actual case the court recognized a discriminatory intent on the part of the employer in enacting the rule, but the analysis would be different. Instead of evaluating whether the employer could show a business necessity for its English-only rule, the court would hold the employer to the stricter BFOQ standard. Since an aspect of the employer's business required employees to speak Spanish to customers on the telephone, the employer could clearly not show that speaking only English was a BFOQ.

B. Changes to the Treatment of Accent Discrimination Claims

The difference that a court's recognition of the full extent of the harm of accent discrimination would make is less clear, since the framework for disparate treatment claims is different than that for disparate impact and hostile work environment claims.[235] However, this understanding should lead courts to inquire more deeply into employers' stated reasons for why they took the accent of an employee into account in making an employment decision. Even though an employer's burden to show a legitimate nondiscriminatory reason is not high, courts should place employers' proffered reasons under some scrutiny.[236] Customers' preferences to not interact with employees who have foreign accents should not outweigh the interest of workers in not having to cover their accents. Thus, customer preference should not be accepted as a legitimate nondiscriminatory reason unless customers truly cannot understand an accent.[237] In recognizing the harm that accent discrimination can do to employees, courts should also more carefully inquire into whether employers' legitimate nondiscriminatory reasons are actually pretexts for discrimination.

Employers should still be able to justify basing employment decisions on accents in some situations. For instance, in *Mejia v. New York Sheraton Hotel*,[238] the plaintiff alleged that she had been denied a promotion from housekeeping to the front desk of a hotel because she was Spanish.[239] Her

235. *See supra* Part I.

236. *See* Tex. Dep't of Cmty. Affairs v. Burdine, 450 U.S. 248, 254–55 (1981) (finding that a defendant need only "raise[] a genuine issue of fact" about the alleged discrimination, but that to do so, it must use admissible evidence to "clearly set forth" a reason for its adverse employment action that would "justify a judgment" in its favor).

237. *See* Bradley v. Pizzaco of Neb., Inc., 7 F.3d 795, 799 (8th Cir. 1993) (rejecting customer preference as a justification for a policy requiring deliverymen to be clean-shaven that had a disparate impact on black men, since being clean-shaven did not relate to how well the deliverymen could do their jobs).

238. 459 F. Supp. 375 (S.D.N.Y. 1978).

239. *Id.* at 376.

employer asserted that she had been denied the promotion because of her very limited English ability.[240] The court scrutinized that reason for itself at trial, and determined that she had an "English language deficiency that made it quite difficult for the Court, the reporter and counsel to understand what she was saying in her testimonial responses."[241] In holding against the plaintiff, the court emphasized the "inability on the plaintiff's part to articulate clearly or coherently."[242] That sort of evaluation—that an employee would simply not be understood by customers—should still be able to justify an employer's accent-based adverse employment decision.

Conclusion

Currently, the jurisprudence surrounding English-only rules and accent discrimination does not do justice to Title VII's prohibition of discrimination on the basis of national origin. Courts generally do not find a significant discriminatory impact in forcing an employee to comply with an English-only rule or passing him over for a job because of his foreign accent, possibly because the judges themselves do not place much value on other languages or manners of speech. These courts fail to appreciate how language and accent are connected to one's sense of national origin identity. Through the framework of forced covering, it may be possible to illustrate to the courts the severity of the harm that English-only rules and accent discrimination cause to the identities of non-native English speakers on the basis of their national origins. If courts appreciate that impact, they should approach English-only rule and accent discrimination lawsuits differently, and more often find that these policies and employment decisions discriminate against workers on the basis of their national origins.

—Braden Beard

240. *Id.*
241. *Id.* at 377.
242. *Id.*

Can Insurgent Courts Be Legitimate Within International Humanitarian Law?*

I. Introduction

Over the past few decades, several armed groups involved in non-international armed conflicts (NIACs) have set up courts and conducted trials.[1] These trials have ranged from prosecutions of individuals for war crimes to civil trials concerning ordinary disputes over land or money.[2] Examples of such courts include those established by the Communist Party of Nepal-Maoist (CPN-M) in Nepal, the Frente Farabundo Martí para la Liberación Nacional (FMLN) in El Salvador, and the Liberation Tigers of Tamil Eelam (LTTE) in Sri Lanka.[3]

International humanitarian law (IHL) enumerates specific rights and obligations of states regarding the passing of sentences in an international armed conflict (IAC). However, IHL contains only two provisions—found in Common Article 3 of the Geneva Conventions[4] (CA3) and Additional Protocol II[5] (AP II)—regarding the passing of sentences in a NIAC. The

---

* I am very grateful to Professor Derek P. Jinks for his advice and guidance in the writing process for this Note and for pushing me to think critically about insurgent courts and their status in international humanitarian law. I am also very grateful to all my mentors in law school—too many to name here—who have made the last three years a challenging but rewarding learning experience. I am especially indebted to Professors Justin Driver and Gretchen S. Sween for teaching me so much about what it means to think like a lawyer, and to Professors James N. Loehlin and James A. Wilson, Jr. for helping me get to law school. I would like to thank everyone on the *Texas Law Review* for their hard work this year, with special thanks to our Managing Editor, Benjamin Shane Morgan. I also thank the Volume 91 Notes Editors—Monica Hughes, Ross MacDonald, Lauren Ross, and Michael Selkirk—for getting this Note ready for publication. All remaining errors are my own. Finally, and importantly, I would like to thank Aie, Baba, Dada, Vahini, the rest of my family, and my close friends for their love, support, and encouragement.

1. Jonathan Somer, *Jungle Justice: Passing Sentence on the Equality of Belligerents in Non-International Armed Conflict*, 89 INT'L REV. RED CROSS 655, 679–82 (2007).

2. Sandesh Sivakumaran, *Courts of Armed Opposition Groups: Fair Trials or Summary Justice?*, 7 J. INT'L CRIM. JUST. 489, 491–95 (2009) (discussing the types of trials that have occurred in insurgent courts in the past); Somer, *supra* note 1, at 680.

3. Sivakumaran, *supra* note 2, at 490.

4. Common Article 3 is the term of art used to refer to the common provision, Article 3, in the Geneva Conventions that regulates conduct of parties in a NIAC. *See, e.g.*, Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field, art. 3, Aug. 12, 1949, 6 U.S.T. 3114, 75 U.N.T.S. 31 [hereinafter Convention I]; Geneva Convention for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Members of Armed Forces at Sea, art. 3, Aug. 12, 1949, 6 U.S.T. 3217, 75 U.N.T.S. 85 [hereinafter Convention II]; Geneva Convention Relative to the Treatment of Prisoners of War, art. 3, Aug. 12, 1949, 6 U.S.T. 3316, 75 U.N.T.S. 135 [hereinafter Convention III]; Geneva Convention Relative to the Protection of Civilian Persons in Time of War, art. 3, Aug. 12, 1949, 6 U.S.T. 3516, 75 U.N.T.S. 287 [hereinafter Convention IV].

5. Additional Protocol II is the term used to refer to one of two treaties meant to supplement the Geneva Conventions that many countries entered into in 1977. Protocol Additional to the Geneva

status of insurgent courts under these provisions is ambiguous. In particular, CA3 does not clearly state whether insurgent courts are legitimate, and, assuming insurgent courts are legitimate, what sorts of procedural protections are expected of such courts. Traditionally, scholars have denied the possibility that armed groups can conduct trials under the IHL framework.[6] And with good reason: there have been several instances of insurgent courts abusing fair trial guarantees and meting out rogue punishment rather than just sentences.[7] For these scholars, a proper interpretation of IHL leads to the conclusion that only a state can conduct trials during a NIAC.[8] Any armed group that establishes courts would be violating IHL, and members of the armed group associated with such courts would be guilty of war crimes.[9]

As armed groups have increasingly resorted to establishing courts and conducting trials, however, other scholars have highlighted a growing need to account for insurgent courts within IHL. This project to account for insurgent courts within IHL leads to three questions: First, is there any interpretation of IHL that would recognize the legitimacy of courts of armed groups? Second, assuming that insurgent courts could be legitimate within IHL, which fair trial guarantees does IHL require of such courts? Third, even if the first two questions can be answered, what types of trials *should* IHL recognize as an appropriate exercise by an armed group?

Two scholars, Sandesh Sivakumaran and Jonathan Somer, have separately proposed interpretations of IHL that answer these questions.[10] Both their solutions succeed to an extent. Their interpretations legitimize insurgent courts within IHL, offer different ways of defining the judicial guarantees IHL requires of such courts, and authorize every type of trial in such courts.[11] However, neither solution is satisfactory. In seeking to

---

Conventions of 12 August 1949, and Relating to the Protection of Victims of Non-International Armed Conflicts (Protocol II), June 8, 1977, art. 6, 1125 U.N.T.S. 609, 613–14 [hereinafter AP II].

6. LIESBETH ZEGVELD, ACCOUNTABILITY OF ARMED OPPOSITION GROUPS IN INTERNATIONAL LAW 68 (2002) ("The general feeling . . . that 'it is difficult to conceive of [IHL] giving insurgents the authority to prosecute and try authors of violations,' thus finds wide recognition in international practice.").

7. *See* Sivakumaran, *supra* note 2, at 491–95 (elaborating on the flaws of the courts established by the CPN-M in Nepal, the FMLN in El Salvador, and the LTTE in Sri Lanka).

8. ZEGVELD, *supra* note 6, at 68.

9. *See id.* at 69 (noting experts' doubts on whether courts created by armed groups could ever fulfill the necessary provisions promulgated in the Geneva Conventions).

10. Sivakumaran, *supra* note 2, at 511–13; Somer, *supra* note 1, at 658. While Sivakumaran and Somer are not the only two scholars to have noted the possibility of accounting for insurgent courts within IHL, they have the most comprehensive examination of the relevant IHL provisions. Other accounts do not concentrate on both the legal basis requirement and the judicial guarantees requirement in IHL. *See, e.g.*, Jan Willms, *Justice Through Armed Groups' Governance – An Oxymoron?* (SFB-Governance Working Paper Series, No. 40, 2012), *available at* http://edocs.fu-berlin.de/docs/servlets/MCRFileNodeServlet/FUDOCS_derivate_000000002183/WP40.pdf?hosts=local.

11. *See* Sivakumaran, *supra* note 2, at 512 ("[A] court that is established by law, conducts fair trials, and contributes to the maintenance of peace and good order among citizens, warrants engagement on the part of the international community."); Somer, *supra* note 1, at 690 (supporting

legitimize insurgent courts within IHL, both scholars ignore undesirable consequences that follow. Specifically, the ultimate goals of both scholars are to improve compliance with IHL and increase humanitarian protection generally during a NIAC.[12] Yet both endorse solutions that lead to a lower level of humanitarian protection. In implicitly arguing for a wholesale loosening of the legal basis requirement in IHL, both scholars ignore the impact such loosening would have on state prosecutions of insurgents. This undercuts the founding impulse that motivates the project to legitimize insurgent courts. Next, in proposing different ways of defining the judicial guarantees IHL requires of insurgent courts, both leave the specific list of required guarantees undefined. Finally, in authorizing every type of trial in an insurgent court, both scholars ignore the substantive differences between different types of trials, and the interpretive difficulties associated with each type.

This Note responds to this discussion of insurgent courts by highlighting some previously ignored interpretive difficulties and argues that any interpretation of IHL that seeks to legitimize insurgent courts leads to problematic solutions. Part II identifies the goals motivating the project to legitimize insurgent courts, discusses why legitimizing insurgent courts within IHL *could* achieve these goals, notes some limiting principles of interpretation that should guide the discussion, and highlights the real dangers posed by insurgent courts. Part III explores the CA3 and AP II provisions governing the passing of sentences in a NIAC. Part IV discusses the legal basis requirement found in CA3 and notes how a loose interpretation of this requirement allows for the existence of insurgent courts. Part V, however, argues against a wholesale loosening of the legal basis requirement because of the impact such a loosening would have on state prosecution of insurgents and relates this discussion to the principle of the equality of belligerents. Part VI examines the fair trial guarantees requirement in CA3, surveys the various methods of defining these guarantees, and proposes a list of guarantees that should apply in a NIAC. Part VII disaggregates the analysis along the dimensions of the type of person to be tried in an insurgent court and the type of trial to occur in such a court, relates this disaggregation to the principle of the equality of belligerents, and argues that any interpretation of IHL that seeks to legitimize

---

the recognition of insurgent courts that endeavor to protect all the "fundamental guarantees" of the judicial system).

12. *See* Sivakumaran, *supra* note 2, at 511–13 ("As well as the very real practical benefits that these courts bring, they provide a means by which to engage the armed group on issues relating to international humanitarian law and the rule of law more generally."); Somer, *supra* note 1, at 690 ("[T]he international engagement of such efforts will not only potentially result in improved compliance with fair trial requirements, but will also create opportunities for broader armed opposition group engagement to encourage compliance with the law of non-international armed conflict in general.").

insurgent courts leads to problematic results. Part VIII offers concluding thoughts.

II. The Motivating Goals of the Legitimization Project and Two Limiting Interpretive Principles

Two interrelated goals motivate the current scholarly drive to legitimize insurgent courts within IHL. First, it is argued that legitimization leads to greater compliance with IHL by armed groups. Second, it is argued that it also increases the general level of humanitarian protection in a NIAC. Sivakumaran and Somer have different explanations about why legitimization achieves the first goal, but only Sivakumaran argues that legitimization achieves the second goal.

Legitimizing insurgent courts promotes compliance with IHL by armed groups, according to Somer, because of the principle of the equality of belligerents.[13] The equality of belligerents is a "fundamental postulate" of *jus in bello* or IHL.[14] It dictates that the rules of IHL should be applied without discrimination to both sides of the conflict.[15] In the arena of an IAC, several reasons necessitate such a principle. First, the lack of an objective method of defining "aggressor" and "aggressed against" means that states are unwilling to accept discriminatory treatment under IHL, but always wish such treatment to be imposed on their enemies.[16] Second, adhering to the equality of belligerents increases humanitarian protection because it does not allow innocent civilians and other protected persons from the wrongdoing state to be treated any differently than similarly situated persons from the wronged state.[17] Third and most importantly, the equality of belligerents ensures compliance with IHL because of the principle of reciprocity.[18] "No belligerent will ever accept that it must apply the rules of warfare against its adversary when this adversary is not itself ready to apply them reciprocally."[19] With the adoption of the equality of belligerents, however, each side feels incentivized to follow the rules if only because the other side also follows the same rules.

While these justifications make sense in the context of an IAC, the equality of belligerents principle may not be a viable concept in a NIAC. To begin, IHL is ambiguous about whether the equality of belligerents is an

13. Somer, *supra* note 1, at 658.

14. YORAM DINSTEIN, THE CONDUCT OF HOSTILITIES UNDER THE LAW OF INTERNATIONAL ARMED CONFLICT 3 (2d ed. 2010) (footnote omitted).

15. ROBERT KOLB & RICHARD HYDE, AN INTRODUCTION TO THE INTERNATIONAL LAW OF ARMED CONFLICTS 23 (2008).

16. *Id.* at 23–25.

17. *Id.* at 25.

18. *Id.*

19. *Id.*

applicable principle in a NIAC.[20] While CA3 "binds each party to the conflict," AP II was almost rejected during negotiations until draft Article 5—which stated that "[t]he rights and duties of the parties to the conflict under [AP II] are equally valid for all [sides]"—and any other provisions that equalized states and armed groups were jettisoned.[21] States are generally hesitant to acknowledge any sort of equality between their governments and an armed group because by definition an armed group seeks to displace the state.[22] Additionally, customary IHL is determined only with reference to state practice, and not with reference to the practice of armed groups.[23] This fact further discredits the idea that the equality of belligerents applies in the relationship between a state and an armed group.[24] Thus, it is tempting to conclude that the principle of the equality of belligerents does not apply in a NIAC.

Somer tries to resurrect the equality of belligerents principle by distinguishing between parity and the equality of belligerents.[25] Parity refers to "a general equality of status as exists between states at international law."[26] Equality of belligerents refers to a narrower concept, one that "does not necessarily mean equal standing, but *equal rights and obligations flowing from the international law norms regulating the subject matter of IHL*."[27] This move to redefine the equality of belligerents narrowly is an important one because states are more likely to accept the principle in a NIAC if it does not imply equality of status. Thus, while the disparity between states and armed groups would justify customary IHL being formed by referring only to state practice, the equality of belligerents would simply grant equal rights and obligations to both sides. Somer believes that the equality of belligerents, as he narrowly defines it, is a necessary principle in a NIAC because it is the only way of ensuring that armed groups feel bound by IHL.[28] The reasoning is simple: if armed groups feel like they have the same rights and obligations under IHL as states, then they are more likely to follow the rules. With respect to the provisions about passing sentences, if the ability of armed groups to conduct trials is not recognized, then the equality of belligerents is denied, and armed groups will have little incentive to comply with IHL. On the other hand, legitimizing insurgent courts would recognize the equality of belligerents, and this would promote compliance

20. Somer, *supra* note 1, at 660.
21. *Id.*
22. *Id.*
23. *Id.* at 661–63.
24. *Id.* at 661–62.
25. *Id.* at 663.
26. *Id.*
27. *Id.*
28. *Id.* at 658 ("An effective principle of equality would require that armed opposition groups have the legal capacity to exercise the rights which flow from the obligations and prohibitions of IHL. Otherwise there is little left to convince them to comply with IHL at all.").

with IHL by armed groups. Therefore, legitimizing insurgent courts would achieve greater compliance with IHL.

Legitimization of insurgent courts would foster compliance with IHL in another way as well. Both Sivakumaran and Somer argue that legitimization would foster compliance because insurgent courts are the only feasible forums that armed groups can use.[29] Armed groups are usually unwilling to transfer their members, accused of having committed violations of IHL, to the state's court system for prosecution.[30] Similarly, transfers of members to the courts of a third-party state or to international criminal courts is also unlikely. The former situation presupposes established relations between the armed group and the third-party state "as well as the consent of all parties involved," while the latter situation does not account for the "jurisdictional constraints" and "limited capacity" of international criminal courts.[31] Therefore, as a matter of practicality, an insurgent court "may be the only forum in which violations of [IHL] will actually be prosecuted."[32] Moreover, Sivakumaran notes that the necessity of such forums is not simply a matter of practicality. Rather, the existence of insurgent courts would help rebel leaders fulfill their command responsibility obligations.[33] In its traditional formulation, as applied to an IAC, command responsibility takes two forms: "[the] responsibility for ordering breaches of international law" and the "responsibility for a subordinate's unlawful conduct that was not directly based on a specific superior order."[34] The trend in IHL has been to extend the concept of command responsibility to leaders of armed groups in NIACs.[35] This trend is reflected in international criminal law, specifically in the Rome Statute of the International Criminal Court, which imposes individual criminal responsibility on rebel leaders for the acts of their subordinates. Under the Rome Statute, rebel leaders are held criminally responsible for the acts of their subordinates where they knew or should have known about the acts but either failed to take "all necessary and reasonable measures" to prevent or repress the acts or did not "submit the matter to competent authorities for investigation and prosecution."[36] Thus, if insurgent courts are the only feasible forums that armed groups can use, such courts would help rebel leaders fulfill their command responsibility obligations.[37] Rebel leaders would be incentivized to refer members of their group, who have committed war crimes, to these insurgent courts in order to avoid

29. Sivakumaran, *supra* note 2, at 510; Somer, *supra* note 1, at 685–86.

30. Sivakumaran, *supra* note 2, at 510.

31. *Id.*

32. *Id.*

33. *Id.*

34. ZEGVELD, *supra* note 6, at 111.

35. *Id.* at 115–17.

36. Rome Statute of the International Criminal Court, art. 28, July 17, 1998, 2187 U.N.T.S. 90 [hereinafter Rome Statute].

37. Sivakumaran, *supra* note 2, at 510; Somer, *supra* note 1, at 685–86.

individual criminal responsibility. As a result, because insurgent courts are the only feasible forums that armed groups could use, legitimizing such courts would promote compliance with IHL.

Besides achieving the first goal of fostering compliance with IHL, Sivakumaran argues that legitimizing insurgent courts also achieves the second goal of increasing the general level of humanitarian protection in a NIAC. Armed groups that hold territorial control often establish courts in addition to providing other services that are usually the domain of the state, like "the provision of education, health services and other manifestations of administrative control."[38] The purpose of providing these services is "to normalize the situation, present the image of a stable, functioning regime and create a quasi-state."[39] Insurgent courts, therefore, promote stability in a NIAC because they "offer an important alternative to summary execution and can contribute to the maintenance of law and order in rebel-held territory."[40] This benefits civilians in two ways. First, insurgent courts act as a check against the possibility of ordinary "criminal gangs flourishing in a climate of impunity."[41] Second, insurgent courts act as a forum in which civilians can bring their claims, even those that do not involve criminal matters. For example, "[m]any cases heard by CPN-M courts involved minor disputes over land, money and familial relationships."[42] Similarly, many of the 23,000 cases heard by LTTE courts involved disputes over land or financial matters.[43] Therefore, insurgent courts increase the general level of humanitarian protection in a NIAC by acting as forums that deal with ordinary criminal and civil matters.[44]

Apart from these two goals that motivate the project to legitimize insurgent courts within IHL, two important interpretive principles must also be followed. First, both Sivakumaran and Somer acknowledge that "[c]ourts of armed opposition groups exist and will continue to exist regardless of the views of third parties."[45] Thus, any interpretation of IHL that legitimizes insurgent courts must do so in a manner that incentivizes armed groups to follow the rules. For example, the solution must both preserve the substance of IHL's fair trial provisions and make compliance by armed groups a real

38. Sivakumaran, *supra* note 2, at 509.

39. *Id.*

40. *Id.* at 490.

41. *Id.* at 509.

42. *Id.* at 492–93.

43. *Id.* at 494.

44. It is important to pause here for a moment. As I will discuss later on, the IHL provisions regarding the passing of sentences govern only penal trials. Thus, IHL does not address civil trials. But since insurgent courts fulfill such a big need by serving as forums for civil disputes, some recognition of their ability to handle such matters must be made. I talk more about this in Part VII.

45. Sivakumaran, *supra* note 2, at 512; *see also* Somer, *supra* note 1, at 690 ("Insurgent courts will continue to operate whether or not they are sanctioned by international law.").

possibility.[46] Any solution that is too idealistic "sacrifices real protection for the sake of paper standards."[47] Moreover, an out-of-reach solution would do nothing to motivate armed groups to conform to the solution, and they would simply keep operating their courts without the blessing of IHL. Thus, a principle of interpretation must be that the solution has to be realistic and one with which armed groups can comply in their day-to-day operations. Second, interpretations that seek to legitimize insurgent courts—like the ones proposed by Sivakumaran and Somer—are arguing against the traditional view that IHL does not leave space for insurgent courts.[48] Thus, any proposed solution can only be justifiable if it raises the level of humanitarian protection in a NIAC. Any solutions that lead to a lower level of protection than would be available under the traditional view should be rejected. These two interpretive principles, therefore, should always guide the search for any solution that legitimizes insurgent courts within IHL.

Before proceeding further, it is important to recognize the potential danger posed by insurgent courts. Both Sivakumaran and Somer acknowledge that insurgent courts often fail to function as forums for fair trials.[49] Indeed, there have been several reports in recent years about the potential for abuse and rogue justice posed by ad hoc courts set up by insurgent groups.[50] The point of the project to legitimize insurgent courts within IHL, however, is not to authorize unfair trials. Nor is it to authorize the existence of all courts that have been established by all insurgent groups. Rather, the attempt to legitimize insurgents courts within IHL seeks to increase humanitarian protection by setting minimum fair trial standards that have to be met.[51] Thus, any insurgent court that fails to meet such standards would be per se illegitimate.

So far, this Note has highlighted the goals behind the drive to legitimize insurgent courts, namely promoting compliance with IHL and increasing the general level of humanitarian protection in a NIAC. Legitimization leads to greater compliance with IHL because of the equality of belligerents principle and because insurgent courts serve as the only feasible forums that can be used by armed groups. Legitimization also increases the general level of protection because insurgent courts can deal with ordinary criminal and civil

---

46. Sivakumaran, *supra* note 2, at 503.

47. *Id.*

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

49. Sivakumaran, *supra* note 2, at 506; Somer, *supra* note 1, at 689.

50. *See* Sivakumaran, *supra* note 2, at 491–95 (noting high profile criticisms by governmental and nongovernmental actors about the insurgent-established FMLN, CPN-M, and LTTE courts, especially regarding the lack of due process guarantees provided by such courts). *See generally* AMERICAS WATCH, VIOLATIONS OF FAIR TRIAL GUARANTEES BY THE FMLN'S AD HOC COURTS (1990) (explaining and condemning the FMLN's ad hoc legal system).

51. Sivakumaran, *supra* note 1, at 512–13 (arguing that the international community should engage in dialogue with the insurgent courts that conduct fair trials and thereby encourage them to enforce international humanitarian law).

matters. Moreover, two limiting interpretive principles have been mentioned: first, the solution must be realistic; second, it cannot lower the level of humanitarian protection than would have been available under the traditional view. Finally, the real dangers posed by insurgent courts have been highlighted. Next, this Note examines the language of the IHL provisions governing the passing of sentences in a NIAC.

III. The CA3 and AP II Provisions Regarding the Passing of Sentences

Two IHL provisions, found in CA3 and AP II respectively, govern the passing of sentences in a NIAC. CA3 was drafted in 1949 as part of the Geneva Conventions, and can be considered to be a microcosm of the Conventions as a whole.[52] AP II was drafted between 1974 and 1977 as part of the diplomatic conference to amend the Conventions.[53] The purpose of AP II, as its first Article proclaims, is to develop and supplement CA3 without modifying CA3.[54] Thus, AP II was meant to inform interpretations of CA3 without changing CA3's content.

In terms of importance in NIACs, CA3 remains much more relevant than AP II for two reasons. Firstly, CA3 applies in more situations than AP II. To begin, the material field of application for CA3 is wider than for AP II. CA3 is meant to apply in all "case[s] of armed conflict not of an international character."[55] AP II, while also applicable in a NIAC, has a higher threshold of application. For AP II to be operative, the armed group must be organized "under responsible command" and must "exercise such control over a part of [the state's] territory as to enable [it] to carry out sustained and concerted military operations and to implement [AP II]."[56] Thus, while CA3 applies at any level of conflict, AP II applies only once a certain threshold has been crossed. Next, more states have signed onto the 1949 Geneva Conventions, and thus CA3, than have signed onto AP II.[57] Thus, CA3 binds more nations than AP II. Furthermore, an armed group is technically incapable of being a party either to the Geneva Conventions or AP II. But since CA3 is now considered customary international law, it binds those parties that have not accepted or are incapable of accepting the Geneva Conventions—including states that have not signed on and armed groups.[58] AP II, however, has not achieved the status of customary international law.[59] This is important for any discussion about an armed

52. *Id.* at 502.

53. *Id.* at 496.

54. AP II, *supra* note 5, art. 1(1).

55. Convention I, *supra* note 4, art. 3.

56. AP II, *supra* note 5, art. 1.

57. One hundred ninety-four countries have signed onto the Geneva Conventions, while only one hundred sixty-six countries have signed onto AP II. *1949 Conventions & Additional Protocols*, INT'L COMM. RED CROSS, http://www.icrc.org/ihl.nsf/CONVPRES?OpenView.

58. Somer, *supra* note 1, at 661.

59. *Id.* at 688.

group's rights and obligations because this means that CA3 is the governing document, not AP II. Secondly, CA3 still applies even when AP II applies because Article 1 of AP II specifically disclaims any intention to modify CA3's "conditions of application."[60] Thus, in AP II conflicts, the provisions regarding passing of sentences from both documents are co-applicable. As a result, because CA3 applies in more situations than AP II and because CA3 applies even when AP II applies, CA3's provision remains the more important one. The CA3 provision governing the passing of sentences is short but full of ambiguity. CA3 prohibits:

> the passing of sentences and the carrying out of executions without previous judgment pronounced by a regularly constituted court affording all the judicial guarantees which are recognized as indispensable by civilized peoples.[61]

Under CA3, a court has to satisfy two conditions. First, it must be "regularly constituted" (also known as the legal basis requirement).[62] Second, it has to afford "all the judicial guarantees which are recognized as indispensable by civilized peoples" (also known as the fair trial guarantees requirement).[63]

While the phrase "regularly constituted" has been used many times in international law treaties, its "precise meaning is less well settled."[64] Similarly, CA3 does not define which judicial guarantees "are recognized as indispensable by civilized peoples." As a result, the CA3 provision provides little meaningful guidance in terms of what type of court could successfully fulfill the legal basis and judicial guarantees requirements.

By contrast, AP II is much more specific about what is required to pass sentences in a NIAC. AP II provides:

> No sentence shall be passed and no penalty shall be executed on a person found guilty of an offence except pursuant to a conviction pronounced by a court offering the essential guarantees of independence and impartiality. In particular:
>
> (*a*) The procedure shall provide for an accused to be informed without delay of the particulars of the offence alleged against him and shall afford the accused before and during his trial all necessary rights and means of defence;
>
> (*b*) No one shall be convicted of an offence except on the basis of individual penal responsibility;

---

60. AP II, *supra* note 5, art. 1.

61. Convention I, *supra* note 4, art. 3(1)(d).

62. Naming this requirement the legal basis requirement is Somer's idea. Somer, *supra* note 1, at 670.

63. Sivakumaran refers to the judicial guarantees as fair trial guarantees. Sivakumaran, *supra* note 2, at 500.

64. *Id.* at 495–96.

> (*c*) No one shall be held guilty of any criminal offence on account of any act or omission which did not constitute a criminal offence, under the law, at the time when it was committed; nor shall a heavier penalty be imposed than that which was applicable at the time when the criminal offence was [committed;] if, after the commission of the offence, provision is made by law for the imposition of a lighter penalty, the offender shall benefit thereby;
>
> (*d*) Anyone charged with an offence is presumed innocent until proved guilty according to law;
>
> (*e*) Anyone charged with an offence shall have the right to be tried in his presence;
>
> (*f*) No one shall be compelled to testify against himself or to confess guilt.[65]

The AP II provision governing the passing of sentences differs from the CA3 provision in two important respects. First, in terms of the legal basis requirement, AP II does away with the CA3 requirement that the court be "regularly constituted."[66] Second, AP II changes the wording of the required judicial guarantees, from those "which are recognized as indispensable by civilized peoples" to those that are "essential guarantees of independence and impartiality."[67] Moreover, AP II clarifies the definition of these guarantees by listing six guarantees that a court must offer in a NIAC.[68]

Both these changes have important implications. By deleting the words "regularly constituted," AP II loosens the legal basis requirement.[69] The deletion was primarily due to the concern of some drafters who felt that armed groups could never fulfill the requirement of having a "regularly constituted" court.[70] Somer notes the apparent discrepancy between AP II's general proclamation that it only develops and supplements but does not modify CA3 and the very substantive modification that occurs in AP II with the change in the legal basis requirement.[71] In any case, AP II seems to leave more room for the establishment of insurgent courts because it does not require they be "regularly constituted," a phrase which as discussed below can be difficult to reconcile with the idea of insurgent courts. Next, by enumerating judicial guarantees, AP II sets a definite standard for what an armed group's courts must achieve in order to ensure a fair trial. Thus, for both Sivakumaran and Somer, these changes make it easier to legitimize

65. AP II, *supra* note 5, art. 6(2).

66. Somer, *supra* note 1, at 670.

67. *Id.*

68. *Id.*

69. *Id.*

70. *Id.*

71. *Id.* at 670–71 ("One may therefore be justified in questioning, in the specific case of the legal basis for the passing of sentences, whether [AP II] which purports to develop [CA3] does not, in fact, end up contradicting it.").

insurgent courts within IHL and specify which judicial guarantees IHL requires of such courts.[72]

Even though these changes resolve two of the three central questions about the interaction between IHL and insurgent courts, their impact is limited. As noted above, the CA3 provision is the more important one because it applies in more situations and is co-applicable with the AP II provision. As a result, any analysis of IHL and its interactions with insurgent courts has to define the terms in CA3. Therefore, this Note will next examine the legal basis requirement in CA3, describe traditional attempts to define it by authorities, and note that a looser interpretation of CA3's legal basis requirement allows for the existence of insurgent courts within IHL.

IV. Loosening CA3's Legal Basis Requirement

CA3 requires that a court in a NIAC be "regularly constituted."[73] This requirement would apply to both state and insurgent courts.[74] However, CA3 does not define "regularly constituted." Traditionally, the "regularly constituted" requirement has been construed as referring to courts established by the state. But both Sivakumaran and Somer argue for a looser interpretation of the legal basis requirement to allow for the existence of insurgent courts within the IHL framework.[75] Indeed, such an interpretation finds support not only in AP II but also in international criminal law.

As a preliminary observation, Sivakumaran notes, "it may be that only the state can conform to the 'regularly constituted' requirement, courts of armed groups being ad hoc in nature."[76] Indeed, several authorities have come to this conclusion in their attempt to define the phrase. The customary IHL study conducted by the International Committee for the Red Cross (ICRC), for example, defines a "regularly constituted" court as a court that "has been established and organized in accordance with the laws and procedures already in force in a country."[77] Such a definition would exclude an insurgent court because by its very nature an insurgent court is not in accordance with the laws and procedures already in force in the country where the rebellion is occurring.

The U.S. Supreme Court has also concluded that a "regularly constituted" court is linked with the state. In *Hamdan v. Rumsfeld*,[78] the Court was faced with the legality of the military commissions established to

---

72. *Id.*; *see also* Sivakumaran, *supra* note 2, at 498 (discussing difficulties of insurgent courts in meeting the "regularly constituted" standard of CA3).

73. Convention I, *supra* note 4, art. 3(1)(d).

74. Sivakumaran, *supra* note 2, at 498.

75. *Id.* at 499–500; Somer, *supra* note 1, at 687.

76. Sivakumaran, *supra* note 2, at 498.

77. *Id.* at 498–99; [1: RULES] JEAN-MARIE HENCKAERTS & LOUISE DOSWALD-BECK, CUSTOMARY INTERNATIONAL HUMANITARIAN LAW 355 (2005).

78. 548 U.S. 557 (2006).

try individuals whom the President had reason to believe were members of al Qaeda or had "engaged or participated in terrorist activities aimed at or harmful to the United States."[79] One of the main issues before the Court was whether the military commissions satisfied the CA3 requirement of a "regularly constituted" court. Justice Alito defined a "regularly constituted" court as one that is simply established in accordance with a state's laws: "I interpret this element to require that the court be appointed or established in accordance with the appointing country's domestic law."[80] Therefore, Justice Alito considered the military commissions to be "regularly constituted."

Justice Stevens's majority opinion, however, went one step further in linking the idea of a "regularly constituted" court with the laws of the state. While acknowledging that "regularly constituted" is not defined by either CA3 or its commentary, the majority opinion referred to the ICRC definition of "regularly constituted" and the commentary to Geneva Convention IV.[81] In particular, the Geneva Convention IV commentary "defines regularly constituted tribunals to include ordinary military courts and definitely exclude[s] all special tribunals."[82] Thus, the majority reasoned that because ordinary military courts in the U.S. are "courts-martial established by congressional statutes," the military commissions were not "regularly constituted."[83] The Court acknowledged that "a military commission can be regularly constituted by the standards of our military justice system only if some practical need explains deviations from court-martial practice."[84] However, since no practical need had been demonstrated, the Court refused to hold anything but courts-martial as "regularly constituted."[85]

Hamdan's holding is significant for the project to legitimize insurgent courts within IHL. In essence, according to the Court's interpretation, the phrase, a "regularly constituted" court, does not simply refer to those courts established by state law. Rather, the phrase refers to those courts that are established by state law and customarily used by the state. Obviously, such a definition would entirely exclude any insurgent court from being "regularly constituted"—not only is an insurgent court not established by state law, but it is never used by the state. These state-centric definitions of "regularly constituted," then, do not permit any space for insurgent courts within CA3.

Yet, for both Sivakumaran and Somer, a compelling reason exists to interpret "regularly constituted" in a way that would allow for the existence of courts of armed groups within CA3. Both scholars cite the interpretation

79. *Id.* at 567–72.
80. *Id.* at 726 (Alito, J., dissenting).
81. *Id.* at 632 (majority opinion).
82. *Id.* (internal quotation marks omitted).
83. *Id.*
84. *Id.* at 632–33 (internal quotation marks omitted).
85. *Id.* at 633.

of James E. Bond, who argued for a loose interpretation of the legal basis requirement.[86] According to Bond, "[t]he regularly constituted court requirement should not be construed too literally. Guerillas, after all, are not apt to carry black robes and white wigs in their back packs."[87] Rather, Bond suggested that the test should be "whether the appropriate authorities, acting under appropriate powers, created the court according to appropriate standards."[88] Extending this line of reasoning, both Sivakumaran and Somer deemphasize the importance of a "regularly constituted" court, and argue that the focus should instead be on whether the insurgent court ensures CA3's judicial guarantees.[89] For Sivakumaran, a loose interpretation of "regularly constituted" would "shift the focus away from the particular manner in which the court is set up and towards the way in which it operates."[90] Similarly, Somer argues that "[a] realistic solution should entail a mixture involving a loose interpretation of the legal basis, with emphasis on the judicial guarantees requirement."[91] Such an interpretation would create space within CA3 for insurgent courts.[92] At the same time, it would also focus attention on what matters most: procedural protections for the accused.

Interestingly, two sources of authority provide support for such an interpretation. First, AP II jettisons the "regularly constituted" requirement.[93] While AP II has limited applicability, because it has a higher threshold of application and has not achieved customary international law status,[94] its loosening of the legal basis requirement is important. This loosening illustrates that the drafters of AP II were concerned about the ability of insurgent courts to fulfill the "regularly constituted" requirement. Indeed, the ICRC commentary to AP II makes clear that the phrase was removed because of the concern of some drafters that armed groups would be unable to establish "regularly constituted" courts.[95] Additionally, it is instructive that the drafters retained the "regularly constituted" requirement in Additional Protocol I (AP I), a treaty that was supposed to amend the Geneva Conventions applying to international armed conflicts.[96] Thus, a loose interpretation of CA3's legal basis requirement could be justified by making reference to the loosening of the legal basis requirement found in AP II. It is

86. Sivakumaran, *supra* note 2, at 499 & n.58; Somer, *supra* note 1, at 673–74 & nn.72–73.

87. James E. Bond, *Application of the Law of War to Internal Conflicts*, 3 GA. INT'L & COMP. L. 345, 372 (1973).

88. *Id.*

89. Sivakumaran, *supra* note 2, at 500; Somer, *supra* note 1, at 687.

90. Sivakumaran, *supra* note 2, at 500.

91. Somer, *supra* note 1, at 687.

92. *Id.*

93. AP II, *supra* note 5, art. 6.

94. *See supra* notes 55–60 and accompanying text.

95. Sivakumaran, *supra* note 2, at 498.

96. Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol 1), art. 75(4), Dec. 7, 1979, 1125 U.N.T.S. 3 [hereinafter AP I].

true, however, that during the discussions for draft Article 10 (which eventually became Article 6), the ICRC delegate made specific reference to the fact that Article 1 of AP II, requiring a high threshold of application, had already been adopted.[97] As Somer notes, such a comment was probably meant to allay the concerns of states about loosening the legal basis requirement.[98] The inference is simple: to get states to agree to a provision that made it easier for insurgent courts to exist within IHL, the states had to be reminded that the provision would only apply in certain situations. While this should give one some pause in loosening the legal basis requirement of CA3, which applies in all NIACs, the project to legitimize insurgent courts requires some interpretive stretching in order to create space for insurgent courts. The point is that AP II does allow for the existence of such courts and that CA3 could also be read in a similar fashion.

Another source of authority also seems to engage in such interpretive stretching. Both Sivakumaran and Somer refer to a similar attempt at redefinition made in the Rome Statute of the International Criminal Court and its accompanying Elements of Crime.[99] The Rome Statute imposes individual criminal responsibility for war crimes.[100] Article 8(2)(c)(iv) defines one specific war crime in a NIAC: "The passing of sentences and the carrying out of executions without previous judgment pronounced by a regularly constituted court, affording all judicial guarantees which are generally recognized as indispensable."[101] While the wording of the Rome Statute is slightly different from the wording of CA3, it "is functionally identical."[102] As a result, to define CA3's "regularly constituted," it makes sense to refer to the definition of the Rome Statute's "regularly constituted." Thus, Somer turns to the Elements of Crime, which were drafted in order to impose criminal responsibility on individuals for breaches of the Rome Statute.[103] Article 8(2)(c)(iv)(4) of the Elements of Crime deals with the "regularly constituted" requirement, and makes it a war crime to pass sentences when:

> There was no previous judgment pronounced by a court, or the court that rendered judgment was not "regularly constituted," *that is, it did not afford the essential guarantees of independence and impartiality*, or the court that rendered judgement did not afford all other judicial

97. Somer, *supra* note 1, at 677.
98. *Id.*
99. Sivakumaran, *supra* note 2, at 499; Somer, *supra* note 1, at 674.
100. Rome Statute, *supra* note 36, art. 8.
101. *Id.* art. 8(2)(c)(iv).
102. Somer, *supra* note 1, at 674.
103. *Id.*

> guarantees generally recognized as indispensable under international law.[104]

Somer notes that this definition confuses the legal basis requirement of CA3 ("regularly constituted") with the fair trial guarantees requirement of AP II ("essential guarantees of independence and impartiality").[105] However, as Sivakumaran argues, this redefinition of "regularly constituted" suggests that a "regularly constituted" court can be interpreted in a looser fashion.[106] Under the Elements of Crime definition, a court fulfills the "regularly constituted" requirement through its fair trial guarantees.[107] Thus, insurgent courts could fulfill CA3's legal basis requirement as long as they assured the judicial guarantees required under CA3. This definition does exactly what Sivakumaran and Somer advocate for: it shifts focus away from the legal basis and towards the judicial guarantees.

The attempt to interpret "regularly constituted" in a looser fashion might seem like it stretches interpretive boundaries, but real authority, in the form of AP II and the Elements of Crime of the Rome Statute, exists for such an interpretation. Such an interpretive move allows for the existence of insurgent courts within IHL by shifting the focus towards the judicial guarantees requirement and away from the legal basis requirement. However, while both Sivakumaran and Somer argue for a wholesale loosening of the legal basis requirement, they ignore an interpretive pitfall that must be avoided. This Note therefore argues against a wholesale loosening of the legal basis requirement.

V. Against a Wholesale Loosening of the Legal Basis Requirement

The redefinition of the legal basis requirement to a looser standard is a necessary predicate for the project to legitimize insurgent courts within IHL. But it also results in the first undesirable consequence of the project which undercuts the goals of fostering compliance with IHL and increasing the general level of humanitarian protection in a NIAC. Moreover, this undesirable consequence violates one of the interpretive principles established earlier because it lowers the amount of humanitarian protection that would have been available under a traditional reading of IHL. The problem involves the principle of the equality of belligerents. As a reminder, the equality of belligerents in the context of a NIAC means "*equal rights and obligations flowing from the international law norms regulating the subject*

---

104. Preparatory Comm'n for the Int'l Criminal Court, *Finalized Draft Text of the Elements of Crime*, art. 8(2)(c)(iv)(4), U.N. Doc. PCNICC/2000/1/Add.2 (2000) [hereinafter *Elements of Crime*] (emphasis added).

105. Somer, *supra* note 1, at 675.

106. Sivakumaran, *supra* note 2, at 499.

107. *Id.*

matter of IHL."[108] In other words, under the equality of belligerents principle, the same rules apply to both the state and the armed group.

Under the principle of the equality of belligerents, the looser interpretation of "regularly constituted" would apply not simply to insurgent courts but also to state courts. Take, for example, the military commissions at issue in *Hamdan*. The Supreme Court ruled that these commissions were illegal, and based some of its reasoning on the fact that the commissions failed the state-centric understanding of "regularly constituted."[109] But under the redefined meaning of the legal basis requirement,[110] the military commissions would pass the "regularly constituted" test because there would no longer be any requirement that the courts be established by state law. Rather, to pass the legal basis test, the commissions would simply have to ensure the judicial guarantees required by CA3. Regardless of one's personal opinion of the U.S. military commissions, this consequence is problematic because, as I demonstrate below, the CA3 judicial guarantees represent only a bare minimum set of protections.[111] Thus, in most cases, CA3 judicial guarantees offer less protection than courts that are "regularly constituted" in the strict sense.

Consider the impact not simply in terms of the military commissions, but more generally. Under a wholesale loosening of the legal basis requirement, states would be free to have a two-tier court system. The first tier would be courts traditionally established by state law, such as civilian courts or courts-martial. The second tier would be courts established to prosecute members of rebel groups. While the first-tier courts would be "regularly constituted" in the strict sense, the second-tier courts would be "regularly constituted" in the loose sense. While the first-tier courts would ensure judicial guarantees as specified by domestic law, the second-tier courts would ensure only CA3 judicial guarantees. Since CA3 judicial guarantees are a bare minimum set of protections, as demonstrated below, the second-tier courts would inevitably ensure fewer procedural protections than the first-tier courts. Thus, while under the traditional view, states would have to prosecute rebels in their normal courts, under the loose interpretation, states would be able to prosecute rebels in special tribunals. In essence, the solutions advocated for by both Sivakumaran and Somer violate one of the limiting principles of interpretation that no solution should be accepted that leads to a lower level of protection. Here, in terms of prosecutions in state courts, rebels would enjoy more protection under the traditional view than under the proposed solutions. Surely, neither Sivakumaran nor Somer would be in favor of such a consequence.

---

108. Somer, *supra* note 1, at 663.

109. *See supra* notes 78–85 and accompanying text.

110. *See supra* notes 86–92 and accompanying text.

111. *See infra* Part VI.

Four responses could be offered in response to this problem. First, one could deny that the looser interpretation of "regularly constituted" applies to states courts, and thus preserve the level of protection afforded by state courts to rebels. Such a solution would acknowledge the asymmetrical relationship between a state and an armed group. However, it would violate the equality of belligerents, a result that is problematic because the whole reasoning behind legitimizing insurgents courts is to recognize the equality of belligerents to ultimately provide incentives for the armed group to follow IHL. It is important to remember that the project to legitimize insurgent courts is an innovative one as most authorities do not view such courts as legitimate.[112] Thus, the role of the equality of belligerents in the project is crucial because it is the driving force for a reading of IHL that allows for insurgent courts. Obviously, it would be unfair to state actors to, on the one hand, use the equality of belligerents to justify insurgent courts, while, on the other hand, deny the equality of belligerents to force states not to lower their judicial protections. Thus, the first response is unavailing. Second, one could accept the equality of belligerents, and, as a consequence, apply the looser definition of "regularly constituted" to state courts. However, as noted above, this response is problematic. Indeed, such an analytical move undercuts the overall goal of achieving better humanitarian protection in a NIAC. Third, one could accept the equality of belligerents, and as a result, reject the whole attempt to loosen the legal basis requirement. However, such an approach would prohibit the legitimacy of insurgent courts as it would read "regularly constituted" strictly.

But a fourth approach offers a pragmatic solution. Under this solution, one would differentiate the interpretation of "regularly constituted" along the axis of the person to be prosecuted. Thus, if either a state court or an insurgent court were attempting to prosecute a member of the opposition force, then a strict interpretation of the legal basis requirement would be applied. As a result, an insurgent court could not legitimately prosecute a member of the state's armed forces. A state court could legitimately prosecute a member of the armed group, but only if such a court were established in accordance with state law. By contrast, if a court were attempting to prosecute a member of its own forces, for example, a loose interpretation of the legal basis requirement would be applied. As a result, an insurgent court could prosecute members of the insurgent group itself. A state court could obviously prosecute members of the state's own armed forces. Moreover, there would be no worry that the state would prosecute such persons in second-tier courts, both for political reasons and also because states are under international human rights law (IHRL) obligations to prosecute such individuals in courts with more judicial guarantees than required by CA3.[113]

112. ZEGVELD, *supra* note 6, at 67–68.

113. *See* Appendix A–D (comparing ICCPR, AP I, and AP II judicial guarantees).

Such a solution would preserve the equality of belligerents and prevent the lowering of standards used in state courts, while allowing room for the existence of insurgent courts that could stabilize rebel territory by hearing prosecutions of rebel forces. Moreover, by legitimizing insurgent courts, this solution would also allow rebel leaders to fulfill their command responsibilities and reduce the general level of impunity in rebel territory. Thus, the main goals of the project to legitimize insurgent courts, that of promoting compliance with IHL and of increasing the general level of humanitarian protection, would be achieved.

So far, this Note has highlighted justifications for interpreting the legal basis requirement in a loose fashion and has offered a solution that authorizes the prosecutions of only certain individuals in order to avoid an important interpretive pitfall in the wholesale loosening of the legal basis requirement. Next, this Note describes some ways of defining the judicial guarantees required by CA3, highlights some problems in such interpretations, and proposes a list of these judicial guarantees.

VI. Judicial Guarantees Under Common Article 3

CA3 requires that a court in a NIAC ensure "all the judicial guarantees which are recognized as indispensable by civilized peoples."[114] As with the legal basis requirement, CA3 does not define these judicial guarantees. But these guarantees could be defined by referring to the fair trial guarantees found in other international norms—such as those in IHRL, IHL for IACs, or AP II.[115] As a general matter, IHRL provides more judicial guarantees than IHL for international armed conflicts, which in turn provides more guarantees than AP II.[116] Good justifications exist for defining CA3's judicial guarantees by referring to one of these other regimes, but each method also has drawbacks. Yet, on the whole, equating CA3's judicial guarantees to those found in AP II provides the most appropriate and pragmatic solution.

Before considering the advantages and disadvantages of each method, two key observations must be made. First, any interpretation of CA3's judicial guarantees provision must be realistic.[117] If the guarantees are

114. Convention I, *supra* note 4, art. 3(1)(d).

115. Sivakumaran, *supra* note 2, at 501.

116. For a complete analysis of which guarantees are common to all three regimes and which are only applicable in some regimes, please refer to Appendix A. I use the ICCPR as representative of IHRL in terms of judicial guarantees. Similarly, I use AP I as representative of IHL for IACs in terms of judicial guarantees.

Essentially, five judicial guarantees are common to the ICCPR, AP I, and AP II. *See* Appendix A. Next, four additional guarantees are found in both the ICCPR and AP I, but not in AP II. *See* Appendix B. However, two guarantees are found in both AP I and AP II, but not in the ICCPR. *See* Appendix C. Finally, each document has certain guarantees that are unique to it. *See* Appendix D.

117. Sivakumaran, *supra* note 2, at 503.

interpreted to include guarantees that simply cannot be met by armed groups, then such an interpretation is meaningless. The whole point of reading "regularly constituted" in a loose fashion was to allow leeway for the existence of insurgent courts. Therefore, reading the judicial guarantees requirement strictly so as to disallow insurgent courts would undercut the project to legitimize insurgent courts within IHL. "Rather, they need to be interpreted in a manner which respects their substance while also making compliance with them possible."[118] Second, a sincere attempt must be made to define and list the judicial guarantees required by CA3. Such an attempt cannot simply involve suggesting that reference be made to the guarantees found in other international norms, while leaving unspecified which guarantees should or should not be ensured by CA3. Instead, justification should be provided for why guarantees in these other international norms ought to be imported into CA3. Sivakumaran argues against such importation because it potentially ignores the relationship between the scope and content of CA3.[119] This worry would hold merit if the guarantees from IHRL or AP I were to be imported into CA3, as both those instruments were designed to apply in very different circumstances. But one should not be too concerned about importing AP II guarantees into CA3. For Sivakumaran, importing AP II guarantees into CA3 fails to recognize the difference between CA3 and AP II conflicts—the fact that some CA3 conflicts do not reach the high threshold of organization, territorial control, and sustained and concerted military operations required of AP II conflicts.[120] However, a CA3 conflict in which an armed group attempts to establish a court will invariably be one that fulfills the threshold requirements of AP II.[121] As a result, importing AP II's guarantees into CA3 should not be problematic. But why should the guarantees in AP II, but not those in IHRL or AP I, be the guiding light? To answer this question, it is important to examine the justifications and drawbacks of each approach.

Under the first method, CA3's judicial guarantees would be interpreted by referring to IHRL's fair trial guarantees.[122] There are several reasons to think that CA3 should offer the same level of protection as IHRL. For instance, while the words "civilized peoples" and "indispensable" are incredibly complex and loaded terms, the thrust of CA3's judicial guarantees requirement is to ensure that courts in a NIAC have at least those guarantees which are universally recognized as necessary for a fair trial.[123] It is not unreasonable to view the IHRL guarantees as representative of such universally recognized rights. Indeed, IHRL is seen as ensuring the most

118. *Id.*
119. *Id.*
120. *Id.*
121. *Id.* at 509–10.
122. *Id.* at 502.
123. Convention I, *supra* note 4, art. 3(1)(d); *see supra* note 63 and accompanying text.

basic human rights.[124] Moreover, the drafting history of AP II suggests that the drafters sought to mimic IHRL's guarantees as codified in the International Covenant on Civil and Political Rights (ICCPR).[125] This is important because in drafting a clarification to and supplement of CA3, delegates were seeking to use the ICCPR to guide their efforts.

But there are several reasons to resist importing the ICCPR guarantees into CA3. First, the ICCPR has a derogation scheme, and derogation applies in times of an internal armed conflict.[126] If the CA3 guarantees are viewed as a bare minimum set of protections that cannot be derogated, and if these CA3 guarantees are defined using the ICCPR guarantees, then that would mean that the ICCPR guarantees are nonderogable. Some commentators have reached this conclusion regarding the nonderogable nature of the ICCPR guarantees and IHRL guarantees generally.[127] However, this is an inelegant solution. Why would the ICCPR framers include a general derogation scheme that applies to the fair trial guarantees if those guarantees are indeed nonderogable? As a result, the fact that the ICCPR guarantees are derogable leads to the conclusion that they cannot be the same as the CA3 guarantees. Second, although the drafters of AP II referred to the ICCPR to guide their attempt to list guarantees, they did not import all the ICCPR guarantees into AP II. Instead, they imported only five ICCPR guarantees into AP II.[128] Moreover, they added one guarantee into AP II that is not found in the ICCPR.[129] This suggests that even the AP II drafters thought that the ICCPR guarantees were not completely suitable in a NIAC. Such an inference militates against importing the ICCPR guarantees into CA3. If the ICCPR guarantees weren't appropriate for AP II conflicts, surely they will not be appropriate for CA3 conflicts. Third, importing ICCPR guarantees into CA3 renders the AP II guarantees wholly redundant. Remember that the set of protections offered by the ICCPR is greater than that offered by AP II.[130] If the CA3 guarantees were equated with the ICCPR guarantees, then a better set of protections would apply under CA3 than under AP II. Thus, even in an AP II conflict, the AP II guarantees would be meaningless because the CA3 guarantees would already be applicable and would provide better protection.

---

124. *See* Sivakumaran, *supra* note 2, at 503 ("[I]nternational instruments relating to human rights offer a basic protection to the human person." (internal quotation marks omitted)).

125. *Id.*

126. International Covenant on Civil and Political Rights, art. 4, Dec. 19, 1966, 999 U.N.T.S. 171 [hereinafter ICCPR]. It is true that Article 4 does not allow derogation of Article 15, which contains some important ICCPR judicial guarantees. *Id.* But there is no prohibition from derogating from Article 14, which contains most of the guarantees. *Id.*

127. *See* Jelena Pejic, *The Protective Scope of Common Article 3: More than Meets the Eye*, 93 INT'L REV. RED CROSS 189, 212 (2011) (noting that certain guarantees, "even though textually derogable, must be considered de facto non-derogable even outside armed conflict").

128. *See infra* Appendix A.

129. *See infra* Appendix D.

130. *See supra* note 116 and accompanying text.

For all these reasons, defining the CA3 guarantees in terms of the guarantees found in IHRL is wholly unsatisfactory.

Under the second option, the CA3 guarantees would be defined in terms of the other fair trial guarantees found in the Geneva Conventions. "[CA3] is no more than a microcosm of the Geneva Conventions as a whole, starting out as it did as a preambular provision to the Civilians Convention, designed to reflect the spirit of that Convention."[131] It would be "only natural" to look at the fair trial guarantees in the Conventions on the whole to interpret CA3's guarantees.[132] Such guarantees are found, for example, in Articles 102 through 108 in Geneva Convention III.[133] Similarly, Articles 66 through 75 in Geneva Convention IV would also be relevant.[134] More importantly, the guarantees found in Article 75 of AP I could be used to interpret CA3's guarantees.[135] Since AP I was drafted after the Conventions, it probably represents the best codification of which guarantees apply during an IAC. Therefore, CA3 guarantees could be defined in reference to AP I guarantees. However, just like importing the ICCPR guarantees, importing AP I guarantees into CA3 would nullify the significance of AP II guarantees. Again, the set of protections offered by AP I is greater than that offered by AP II. Thus, if CA3 guarantees were equated with AP I guarantees, a better set of protections would apply under CA3 than under AP II. Thus, even in an AP II conflict, the AP II guarantees would be meaningless because the CA3 guarantees would already be applicable and would provide better protection. As a result, defining CA3 guarantees with reference to AP I guarantees is not a good solution either.

Under the last approach, the CA3 judicial guarantees would be equated with the AP II guarantees. This approach is intuitive because the drafters of AP II understood it to be a supplement to CA3.[136] Indeed, "[i]n introducing what was to become Article 6, the delegate of the ICRC, at the diplomatic conference of 1974–1977, made clear the link between that provision and [CA3], going so far as to run the two together."[137] Moreover, as discussed above, the Elements of Crime of the Rome Statute defines the wording it borrows from CA3 by referring to AP II.[138] Finally, AP II was drafted to apply in the context of a NIAC, albeit one of a higher threshold.[139] However,

131. Sivakumaran, *supra* note 2, at 502.

132. *Id.*

133. Convention III, *supra* note 4, art. 102–108.

134. Convention IV, *supra* note 4, art. 66–75.

135. AP I, *supra* note 96, art. 75. Sivakumaran notes that such an approach was used in Justice Stevens's opinion in *Hamdan*. Sivakumaran, *supra* note 2, at 502; *see also* Hamdan v. Rumsfeld, 548 U.S. 557, 633 (2006).

136. Sivakumaran, *supra* note 2, at 501–02.

137. *Id.* at 501.

138. *See supra* notes 101–06 and accompanying text.

139. *See supra* notes 55–56 and accompanying text.

since AP II was drafted for a similar context, it is reasonable to look to AP II. As a result, importing the AP II guarantees seems like the best option.

Even this approach, though, has challenges that must be accounted for. First, such an approach would impose AP II guarantees on CA3 conflicts. Thus, if the NIAC was to occur in a state that refused to ratify AP II, this approach imposes AP II on the state despite the state's refusal to ratify the treaty. This has serious implications in terms of state sovereignty. Second, there is some evidence that when the AP II drafters wrote the fair trial provision, they were keenly aware of its applicability only at a higher threshold. For example, at the negotiations where the AP II provision was discussed, "[t]he ICRC delegate began the discussion by emphasizing that draft Article 10 [which eventually became Article 6] should be considered in light of the fact that Article 1 on the high threshold of application, including territorial control, had already been passed by the drafting committee."[140] As Somer notes: "The intention of such a comment was most probably to ensure that states recognized that the adoption of a provision with a wider scope of application than [CA3] would only be applicable to high-threshold conflicts."[141] Thus, it could be argued that states would not want AP II guarantees applying in a CA3 conflict.

However, these issues are less problematic than they appear. As to the first objection, it is important to keep in mind that any sort of attempt to define CA3's guarantees will inevitably involve importation from another regime. Any such importation—either from the ICCPR, AP I, or AP II—will result in the imposition of standards that a state has not necessarily agreed to. For example, if a NIAC is occurring in a country that has refused to sign the ICCPR, interpreting CA3's guarantees as being equivalent to those in the ICCPR would also raise similar sovereignty concerns. Therefore, this problem is non-unique and a necessary by-product of any attempt to actually define CA3's guarantees. As to the second objection, it should be kept in mind that the concern of the delegates was about the loosening of the legal basis requirement in AP II. Thus, to allay the fears of the delegates, the ICRC representative reminded them of the high threshold requirement of AP II. However, this does not necessarily preclude importing AP II guarantees into CA3. Indeed, such an importation would be least harmful to the relationship between the scope and content of CA3. Although AP II was designed to apply in a conflict of a higher threshold, its context of application (NIACs) is still much closer to the context of CA3 than it would be to the context of the ICCPR or AP I. Therefore, CA3's judicial guarantees requirement ought to be interpreted as being the same as those in AP II.

Two final issues must be addressed. The first issue deals with the potential convergence of international norms regarding fair trial guarantees.

140. Somer, *supra* note 1, at 677.

141. *Id.*

Sivakumaran argues that in reality the provisions in IHRL, AP I, and AP II all converge on the same guarantees.[142] Both AP I and AP II require the court to "afford the accused before and during his trial all necessary rights and means of defence."[143] Sivakumaran finds this phrase ambiguous, and therefore turns to the only international drafting committee that has attempted to define it. In drafting the Elements of Crimes for the Rome Statute, the drafting committee had to specify what the phrase "all necessary rights and means of defence" meant.[144] One state proposed a list of guarantees that could define that phrase.[145] Even though the list was ultimately not adopted for reasons unrelated to its acceptability, the list is striking because it seems to incorporate into AP I and AP II all the ICCPR guarantees that are missing.[146] For Sivakumaran, the fact that the proposed list aligns the AP I, AP II, and ICCPR guarantees demonstrates that international norms actually converge on the question of which guarantees should be ensured in a fair trial.[147] However, this argument of convergence should be strenuously resisted. For one thing, the list of proposed guarantees was never adopted.[148] More importantly, it should always be remembered that AP I and AP II were drafted after the ICCPR.[149] Therefore, any fair trial guarantees in the ICCPR that were left out of either treaty must have been intentionally left out as unsuitable to IACs or NIACs. Finally, attempting to align the guarantees in all three documents would make it practically impossible for an insurgent court to fulfill the fair trial guarantees requirement as the burden on the insurgent court would be too high. Thus, the conclusion has to be that there is no convergence: the ICCPR offers more guarantees than AP I, which in turn offers more guarantees than AP II.

The second issue deals with the use of the word "law" in the AP II guarantees. Specifically, the word "law" is found in Article 6(2)(c) and 6(2)(d).[150] During the drafting of the AP II guarantees, when a draft of the provision used the expression "national or international law," instead of simply "law," states objected that such language allowed for the possibility of insurgent law.[151] The worry was that courts during NIACs could apply

142. Sivakumaran, *supra* note 2, at 505.

143. AP I, *supra* note 96, art. 75(4)(a); AP II, *supra* note 5, art. 6(2)(a).

144. Sivakumaran, *supra* note 2, at 504.

145. *Id.*

146. *Id.* at 504–05.

147. *Id.* at 505.

148. *See id.* (describing three separate reasons why the proposed list of fair trial guarantees was disfavored); Knut Dörmann, *War Crimes Under the Rome Statute of the International Criminal Court, with a Special Focus on the Negotiations on the Elements of Crimes*, 7 MAX PLANCK Y.B. UNITED NATIONS L. 341, 399 (2003) ("Instead of weakening the value of such a list of fair trial guarantees by an introductory paragraph defining what is to be considered indispensable, states preferred not to include such a list.").

149. *See supra* notes 5, 96, 126 and accompanying text.

150. AP II, *supra* note 5, art. 6(2)(c)–(d).

151. Somer, *supra* note 1, at 678.

some other law besides the state's law.[152] This is problematic because if AP II guarantees are imported into CA3, and the word "law" refers only to the state's law, then insurgent courts would be bound to apply the law of the opposing side. As Sivakumaran notes, such a solution would be nonsensical, and would hardly gain acceptance by armed groups.[153] As a result, in understanding the word "law" in the AP II guarantees, one should interpret the provision as referring to the law of the party in whose court the accused is being tried. This would allow for insurgent courts to apply insurgent laws.

In conclusion, the judicial guarantees in CA3 should be interpreted as being the same as those in AP II. Even though this approach has its drawbacks, it presents the most pragmatic solution to the problem and allows interpreters to come up with a definite list of guarantees that should apply in CA3 conflicts. It also preserves as much as possible the connection between the scope and content of CA3. There is no perfect solution to the dilemma of defining the judicial guarantees in CA3. However, one must not give up the task entirely because left undefined the CA3 guarantees offer almost no guidance on what is and is not acceptable.

In analyzing insurgent courts, this Note has so far recounted the arguments put forth by both Sivakumaran and Somer about loosening the legal basis requirement, has argued against such a wholesale loosening, and has defined the judicial guarantees required by CA3. Next, it will disaggregate the analysis of insurgent courts along the axis of the type of person to be tried and the type of trial, and will argue that any interpretation of IHL that legitimizes insurgent courts leads to problematic results.

VII. What Types of Trials Should Be Legitimized?

Even though a loose interpretation of the legal basis requirement creates space for insurgent courts within IHL, and even though an insurgent court could possibly meet the judicial guarantees found in CA3, a *policy* judgment must still be made. What types of trials *should* be recognized as legitimate within IHL? This question can only be answered by disaggregating the analysis along two axes: the type of person to be prosecuted in an insurgent court and the type of prosecution. However, any policy judgment—in other words, any interpretation of IHL that recognizes the legitimacy of insurgent courts—leads to problematic results.

To begin, there are three types of people who could be prosecuted in an insurgent court: members of the state's armed forces, members of the armed group itself, and civilians. Remember, as argued before, the equality of belligerents has already put members of the state's armed forces out of the reach of insurgent courts. Thus, that leaves only two categories of individuals whom insurgent courts could possibly prosecute.

152. *Id.*

153. Sivakumaran, *supra* note 2, at 508.

Next, there are four kinds of trials an insurgent court could hear: (1) prosecutions for mere participation in hostilities against the armed group;[154] (2) prosecutions for war crimes or for violations of IHL;[155] (3) prosecutions for violations of the insurgent group's penal code covering ordinary crimes;[156] and (4) trials involving civil disputes.[157]

Before continuing the analysis, special notice must be given to the fact that IHL does not actually govern the ability of armed groups to hear civil disputes. For example, CA3 prohibits the "passing of sentences and the carrying out of executions" unless a judgment has been rendered by a court fulfilling the legal basis and judicial guarantees requirements.[158] Clearly, CA3's provision only relates to penal prosecutions because "sentences" are traditionally understood to be adjudications in a criminal context.[159] AP I and AP II make the connection between their provisions and penal prosecutions clearer. AP I, for example, specifies that its provision is applicable only in the context of a "penal offence related to the armed conflict."[160] AP II specifically relates its provision "to the prosecution and punishment of criminal offences related to the armed conflict."[161] Thus, none of the IHL provisions either authorize or deny armed groups' ability to adjudicate civil disputes. Yet there is some evidence that insurgent courts serve an important function as forums for civil disputes.[162] As a result, in the project to legitimize insurgent courts within IHL, their authority to resolve civil disputes should be discussed.

Turning back to the initial question about the extent to which insurgent courts can pass sentences or resolve civil disputes involving either members of the armed group or civilians, it becomes apparent that any interpretation that legitimizes insurgent courts leads to a problematic solution. For example, let's begin with the most palatable solution under which an insurgent court could be legitimate under IHL. Under such a solution, an armed group's ability to prosecute individuals is the most constrained.

---

154. Somer, *supra* note 1, at 683.

155. *Id.* at 682.

156. While neither Sivakumaran nor Somer deal with this category of trials, what I have in mind are those prosecutions for crimes that are not committed in relation to the armed conflict, such as one civilian murdering another for pecuniary profit. Such a prosecution would usually be handled under domestic criminal law.

157. *See, e.g.*, Sivakumaran, *supra* note 2, at 492–93 ("Many cases heard by CPN-M courts involved minor disputes over land, money and familial relationships.").

158. Convention I, *supra* note 4, art. (3)(1)(d).

159. *See* BLACK'S LAW DICTIONARY 1485 (9th ed. 2009) (defining "sentence" as "[t]he judgment that a court formally pronounces after finding a criminal defendant guilty; the punishment imposed on a criminal wrongdoer").

160. AP I, *supra* note 96, art. 75(4).

161. AP II, *supra* note 5, art. 6(1).

162. *Cf.* Sivakumaran, *supra* note 2, at 512 (stressing that certain real and practical benefits of such courts must be considered in evaluating the proper level of engagement for the international community).

Under such a solution, in terms of members of the armed group's own forces, insurgent courts would be capable of hearing both penal prosecutions of such individuals as well as civil disputes in which such individuals were involved. Arguably, recognizing the legitimacy of both types of trials involving individual insurgents reduces the level of chaos in rebel territory and the feeling of impunity among members of the rebel group. It allows the insurgent court to fulfill the need for a feasible forum that armed groups could realistically use. More importantly, recognizing the legitimacy of penal prosecutions of such individuals, especially prosecutions for war crimes and ordinary crimes, allows rebel leaders to fulfill their command responsibility obligations. However, one might question whether insurgent courts should be allowed to prosecute the armed group's own members for mere participation offenses. Such a situation would probably involve a rebel group member who turned on the group and aided the state in some fashion. The issue in such a situation would be whether such a person deserves greater protection—on level with members of the state's armed forces. Therefore, under the most constrained solution, an insurgent court would not be allowed to prosecute the armed group's own members for mere participation offenses.

Next, under the most constrained solution, in terms of civilians, the insurgent court would not be allowed to hear either penal prosecutions or civil disputes. Such a limitation would be justified because of the concerns of abuse and rogue punishment. Both Sivakumaran and Somer respond to such concerns by simply stating that even state courts have been known to violate fair trial guarantees.[163] However, such a response does not take into account the very real concerns about the dangers of insurgent courts.[164]

Therefore, under the most constrained solution, an insurgent court's legitimacy within IHL would look like this:

163. Sivakumaran, *supra* note 2, at 506; Somer, *supra* note 1, at 690.

164. *See supra* Part II.

Most Constrained Solution: Trials by Insurgent Courts That Should Be Legitimized Under IHL				
		Types of Individuals		
		I. Members of the State's Armed Forces	II. Members of the Armed Group	III. Civilians
Types of Trials	A. Prosecutions for Mere Participation Offenses	NO	NO	NO
	B. Prosecutions for War Crimes	NO	YES	NO
	C. Prosecutions for Ordinary Penal Crimes	NO	YES	NO
	D. Civil Disputes	NO	YES	NO

Such a constrained solution is the most palatable because it constrains an insurgent court the most—and thus checks for abuses by such courts. However, it raises problems in terms of the equality of belligerents. Essentially, the justification for such a solution would be to argue that the reason an insurgent court does not have the legitimacy to prosecute civilians categorically and certain members of the armed group's own forces is because it cannot establish regularly constituted courts in the strict sense. Thus, while a state can fulfill that requirement by simply using a state court, the armed group cannot fulfill such a requirement. Furthermore, the justification would go, the only situation in which a loose interpretation of "regularly constituted" is applied is in the three exceptions in the chart above. Technically, such a solution would preserve the equality of belligerents, while also legitimizing insurgent courts as narrowly as possible in order to allow them to serve as a way for rebel leaders to fulfill their command responsibility of policing their groups. While this solution is technically sound, it is unlikely to be seen as unbiased by insurgent groups, and therefore unlikely to be followed. As a result, such an interpretation results in a problematic solution.

On the other hand, IHL could be interpreted much more broadly. But even a broader interpretation would be problematic. Under the broad

solution, insurgent courts would still not be allowed to prosecute members of the state's armed forces due to the equality of belligerents, as discussed above. However, in terms of the armed group's own members, the insurgent court would be able to hear prosecutions and civil trials without any qualification. Thus, an insurgent court would be able to hear prosecutions of its own members for mere participation crimes. It could be argued that rebels facing such mere participation prosecutions are likely to benefit from the political connections they possess with the rebel group. This lessens any concerns over abuse by the insurgent court. In terms of civilians, under the broad solution, insurgent courts would be free to hear prosecutions and civil trials without qualification as well. Insurgent courts benefit the civilian population in two ways. First, they reduce the "climate of impunity" by working as a counterbalance against ordinary criminal gangs.[165] Second, they also serve as forums in which civilians can seek redress for their civil disputes.

Therefore, under the most broad solution, an insurgent court's legitimacy within IHL would look like this:

Most Broad Solution: *Trials by Insurgent Courts That Should Be Legitimized Under IHL*				
		Types of Individuals		
		I. Members of the State's Armed Forces	II. Members of the Armed Group	III. Civilians
Types of Trials	A. Prosecutions for Mere Participation Offenses	NO	YES	YES
	B. Prosecutions for War Crimes	NO	YES	YES
	C. Prosecutions for Ordinary Penal Crimes	NO	YES	YES
	D. Civil Disputes	NO	YES	YES

165. Sivakumaran, *supra* note 2, at 509.

This solution is also highly problematic. First, it leaves civilians, a highly vulnerable population during war, in the hands of insurgent courts. As has been discussed above, insurgent courts are often associated with abuse of fair trial guarantees, and do not pass sentences based on just principles but often due to political reasons. Moreover, unlike members of the armed group, civilians are less likely to possess political connections within the group. Thus, they will face an added danger in prosecutions for any kind of offense. Even in terms of civil disputes, although civilians would not be at risk of penal punishment, insurgent courts could use civil fines or remedies as a means of abusing the civilian population. Second, it leaves rebels who have decided to switch sides, another vulnerable population, in the hands of insurgent courts as well. While this population will enjoy more political connections than civilians, they are still at risk of facing an unfair trial. Obviously, therefore, the most broad solution does not provide a satisfying answer either. Indeed, it is more troublesome than the most constrained solution.

The point here is that any interpretation of IHL that legitimizes insurgent courts produces problematic solutions. While the most constrained solution is problematic because it is less likely to be accepted as unbiased by insurgent groups, the most broad solution is problematic because of the concerns of abuse and violation of fair trial guarantees. Although this does not mean that there is no solution categorically, it means that further discussion is needed for what ultimately is a policy judgment. At the end of the day, the real question is less a question of law than of policy: how much is the international community willing to risk in order to legitimize (and thus, hopefully engage) insurgent courts? Thus, while both Sivakumaran and Somer might be right that insurgent courts can be legitimized under IHL, the question left unanswered is to what extent and at what price? The point of this Note is to highlight the complexities involved in answering that question, and to suggest that any solution adopted will be problematic in one way or another.

VIII. Conclusion

Trials by armed groups have the possibility of promoting compliance with IHL and increasing the general level of humanitarian protection in a war-torn country. However, in interpreting IHL, several difficulties must be dealt with. First, in terms of the equality of belligerents principle, it is clear that the legitimacy of insurgent courts to prosecute members of the state's armed forces must be denied. Next, one must also be careful in legitimizing proceedings that involve either members of the armed group itself or civilians. No matter what solution is adopted, there will be problematic consequences. Ultimately, the international community has to decide the question as a matter of policy, and not simply as a matter of international law.

The ultimate goal is to offer a pragmatic solution that both increases protection and incentivizes armed groups to follow the rules. Anything too idealistic, though, would ignore the reality that insurgent courts largely operate outside the boundaries of IHL. The purpose of this Note has been to argue that such courts can be properly accounted for within IHL, but that any interpretation that does so leads to a problematic solution.

—Parth S. Gejji

Appendix A				
Fair Trial Guarantees Found in All Three International Norms				
		ICCPR	AP I	AP II
Fair Trial Guarantee	Right to be presumed innocent until proven guilty.	Article 14(2) ("Everyone charged with a criminal offence shall have the right to be presumed innocent until proved guilty according to law.")	Article 75(4)(d) ("[A]nyone charged with an offence is presumed innocent until proved guilty according to law.")	Article 6(2)(d) ("[A]nyone charged with an offence is presumed innocent until proved guilty according to law.")
	Right to be promptly informed of the charge.	Article 14(3)(a) ("To be informed promptly and in detail in a language which he understands of the nature and cause of the charge against him.")	Article 75(4)(a) ("[T]he procedure shall provide for an accused to be informed without delay of the particulars of the offence alleged against him . . . .")	Article 6(2)(a) ("[T]he procedure shall provide for an accused to be informed without delay of the particulars of the offence alleged against him . . . .")
	Right to be tried in presence.	Article 14(3)(d) ("To be tried in his presence . . . .")	Article 75(4)(e) ("[A]nyone charged with an offence shall have the right to be tried in his presence.")	Article 6(2)(e) ("[A]nyone charged with an offence shall have the right to be tried in his presence.")
	Right not to testify or confess guilt.	Article 14(3)(g) ("Not to be compelled to testify against himself or to confess guilt.")	Article 75(4)(f) ("[N]o one shall be compelled to testify against himself or to confess guilt.")	Article 6(2)(f) ("[N]o one shall be compelled to testify against himself or to confess guilt.")
	Right not to be held guilty unless the act or omission constituted a criminal offense at the time of its occurrence. Right not to have a heavier penalty imposed than was the sentence at the time of the act or omission. Right to a lesser penalty if a lesser penalty is subsequently enacted into criminal law for that act or offense.	Article 15(1) ("No one shall be held guilty of any criminal offence on account of any act or omission which did not constitute a criminal offence, under national or international law, at the time when it was committed. Nor shall a heavier penalty be imposed than the one that was applicable at the time when the criminal offence was committed. If, subsequent to the commission of the offence, provision is made by law for the imposition of the lighter penalty, the offender shall benefit thereby.")	Article 75(4)(c) ("[N]o one shall be accused or convicted of a criminal offence on account of any act or omission which did not constitute a criminal offence under the national or international law to which he was subject at the time when it was committed; nor shall a heavier penalty be imposed than that which was applicable at the time when the criminal offence was committed; if, after the commission of the offence, provision is made by law for the imposition of a lighter penalty, the offender shall benefit thereby.")	Article 6(2)(c) ("[N]o one shall be held guilty of any criminal offence on account of any act or omission which did not constitute a criminal offence, under the law, at the time when it was committed; nor shall a heavier penalty be imposed than that which was applicable at the time when the criminal offence was committed; if, after the commission of the offence, provision is made by law for the imposition of a lighter penalty, the offender shall benefit thereby.")

Appendix B				
Fair Trial Guarantees Found Only in ICCPR and AP I				
		ICCPR	AP I	AP II
Fair Trial Guarantee	Right to have the judgment publicly pronounced.	Article 14(1) ("[A]ny judgement rendered in a criminal case or in a suit at law shall be made public except where the interest of juvenile persons otherwise requires or the proceedings concern matrimonial disputes or the guardianship of children.")	Article 75(4)(i) ("[A]nyone prosecuted for an offence shall have the right to have the judgement pronounced publicly.")	Not Available
	Right to examine witnesses or have witnesses examined. Right to have witnesses both for and against examined under the same conditions.	Article 14(3)(e) ("To examine, or have examined, the witnesses against him and to obtain the attendance and examination of witnesses on his behalf under the same conditions as witnesses against him.")	Article 75(4)(g) ("[A]nyone charged with an offence shall have the right to examine, or have examined, the witnesses against him and to obtain the attendance and examination of witnesses on his behalf under the same conditions as witnesses against him.")	Not Available
	Right to no double jeopardy in the same jurisdiction.	Article 14(7) ("No one shall be liable to be tried or punished again for an offence for which he has already been finally convicted or acquitted in accordance with the law and penal procedure of each country.")	Article 75(4)(h) ("[N]o one shall be prosecuted or punished by the same Party for an offence in respect of which a final judgement acquitting or convicting that person has been previously pronounced under the same law and judicial procedure.")	Not Available
	Right to be informed of the charge in a language that the accused understands.	Article 14(3)(a) ("To be informed promptly and in detail in a language which he understands of the nature and cause of the charge against him.")	Article 75(3) ("Any person arrested, detained or interned for actions related to the armed conflict shall be informed promptly, in a language he understands, of the reasons why these measures have been taken.")	Not Available

Appendix C				
Fair Trial Guarantees Found Only in AP I and AP II				
		ICCPR	AP I	AP II
Fair Trial Guarantee	Right to be convicted only on the basis of individual penal responsibility.	Not Available	Article 75(4)(b) ("[N]o one shall be convicted of an offence except on the basis of individual penal responsibility.")	Article 6(2)(b) ("[N]o one shall be convicted of an offence except on the basis of individual penal responsibility.")
	Right to only be advised of available judicial remedies upon conviction.	Not Available	Article 75(4)(j) ("[A] convicted person shall be advised on conviction of his judicial and other remedies and of the time-limits within which they may be exercised.")	Article 6(3) ("A convicted person shall be advised on conviction of his judicial and other remedies and of the time-limits within which they may be exercised.")

Appendix D

Fair Trial Guarantees Unique to Each Document

ICCPR		AP I		AP II	
Right of all parties to be equal before the court.	Article 14(1)	Right to be released from detention (except for detention related to penal offenses).	Article 75(3)	Right of persons younger than 18 years to not have the death penalty pronounced. Right of pregnant women and mothers of young children to not have death penalty carried out.	Article 6(4)
Right to a fair and public hearing, with some exceptions.	Article 14(1)	Right of women to be held in separate quarters from men. Right of women to be held in the same place as their families.	Article 75(5)		
Right to have adequate time and facilities for the preparation of defense and to communicate with counsel of one's choosing.	Article 14(3)(b)	Right to be protected by AP I until final release, repatriation, or reestablishment.	Article 75(6)		
Right to be tried without undue delay.	Article 14(3)(c)				
Right to defend self in person or through counsel. Right to have counsel. Right to have counsel paid for.	Article 14(3)(d)				
Right to have a free interpreter in court.	Article 14(3)(f)				
Right of juveniles to different procedures.	Article 14(4)				
Right of appeal, mandatory.	Article 14(5)				
Right to compensation for miscarriage of justice.	Article 14(6)				

* * *

Voluntary Incentive Auctions and the Benefits of Full Relinquishment*

I. Introduction

In February 2012, Congress passed the Middle Class Tax Relief and Job Creation Act, authorizing the Federal Communications Commission (FCC) to use voluntary incentive auctions to repurpose electromagnetic spectrum.[1] These auctions give television broadcasters, to whom spectrum is currently allocated, the option to voluntarily sell their spectrum back to the government. The relinquished spectrum can then be relicensed for multiple uses and reauctioned to companies that supply mobile data plans. To participate, broadcasters must either (1) relocate from their current channel to

* I am grateful to Professor Jane M. Cohen and Professor Matthew L. Spitzer for providing inspiration for this Note and for their thoughtful and careful feedback throughout the writing and editing process. I also want to thank the Volume 91 Notes editors, Monica Hughes, Ross MacDonald, Lauren Ross, and our Editor in Chief, Parth Gejji, for their tenacity and selflessness in preparing this Note for publication, but most importantly, for their friendship. Finally, I would like to thank my mom, my dad, and my sister, Kaethe, for their patience, encouragement, and unconditional love.

1. Middle Class Tax Relief and Job Creation Act of 2012, Pub. L. No. 112-96, § 6402, 126 Stat. 156, 224 (to be codified at 47 U.S.C. § 309(j)(8)(G)).

a shared channel, (2) transfer to a new frequency, or (3) fully relinquish their spectrum.

While the statutory scheme does not unilaterally revoke broadcasting licenses, its structure limits the FCC's ability to redistribute spectrum from its relatively low-value television use to a more high-value wireless broadband use. This Note argues that, from policy and economic perspectives, full relinquishment best accomplishes the goal of optimal spectrum reallocation. Congress likely assumed fully relinquishing broadcasters would be forced off the air after losing access to the spectrum on which their programming was transmitted. This assumption, however, is faulty. Broadcasters choosing to fully relinquish can continue to transmit using other mediums, thereby providing consumers with the services previously offered over the airwaves.

Accordingly, Congress or the FCC should create incentives to encourage broadcasters to choose the full-relinquishment option and protect viewers from potential programming losses. Motivating broadcasters to choose full relinquishment can be accomplished by extending incentives that are already being offered to broadcasters choosing the other two options: sharing or relocation. To incentivize full relinquishment, Congress or the FCC can extend must-carry privileges, pay for broadcasters to relocate to leased-access cable, or include a financial premium.

Part II of this Note discusses the modern regulatory framework, focusing on the inadequate mechanisms for repurposing spectrum and introduces the voluntary incentive auction. Parts III and IV provide an overview of the Middle Class Tax Relief and Job Creation Act, and argue that full relinquishment is the optimal broadcaster choice to help curb the spectrum deficiencies. Part V outlines possible incentives to coax broadcasters to fully relinquish and transmit on other mediums. Part VI briefly concludes.

II. Command and Control in the Modern Era

A. Overview of Modern Command-and-Control Regulations

Despite its critics, FCC stewardship of the electromagnetic spectrum has not changed in a meaningful way since Congress first cleared the airwaves with the Radio Act of 1927.[2] Pursuant to its authority to grant licenses according to "public interest, convenience, and necessity"[3] the FCC has par-

2. THOMAS G. KRATTENMAKER & LUCAS A. POWE, JR., REGULATING BROADCAST PROGRAMMING 12 (1994) (citing the Radio Act of 1927, Pub. L. No. 69-632, 44 Stat. 1162 (repealed 1934 and 1966)). The Radio Act of 1927 was passed with the purpose of eliminating harmful interference caused by overlapping signals, which disrupt and distort the original transmission, leaving the end user unable to comprehend it. STUART MINOR BENJAMIN ET AL., TELECOMMUNICATIONS LAW AND POLICY 55 (3d ed. 2012).

3. 47 U.S.C. § 309(a) (2006).

celed out spectrum for use as a public resource since its creation in 1934.[4] Because frequencies cannot be occupied by two transmissions at the same time without harmful interference, and there are a limited number of frequencies, the government treated spectrum as uniquely scarce.[5] Accordingly, the government retained control over spectrum while vesting ownership in the public.

In the current scheme, known as command and control because of the FCC's direct regulation, spectrum is allocated and licensed for a particular use for a term of up to eight years.[6] Once allocated, the FCC grants a license to a specific individual, organization, or corporation to transmit over a specific frequency, at a specific location, at a specific time, subject to certain parameters of service.[7] Licenses may be renewed subject to public interest considerations, but most licenses are renewed automatically.[8] In this way, the FCC is able to tether frequencies to certain technologies indefinitely.

Consistent with the Radio Act's categorical declaration that there would be no private property interest in spectrum,[9] the rights and privileges of licensees are limited. For example, licensees may not operate after the license expires,[10] subdivide spectrum rights to be transferred,[11] or transfer or reassign the license without FCC approval.[12] Furthermore, as a condition of being granted a free license to broadcast, television stations agree to certain programming restrictions that the FCC believes serve the public interest.[13]

---

4. 47 U.S.C § 301; Red Lion Broad. Co. v. FCC, 395 U.S. 367, 399 & n.26 (1969). For example, the FCC has licensed spectrum to be used for aviation, public safety, FM and AM radio, television broadcasting, and amateur uses, to name a few. *Spectrum Dashboard*, FED. COMM. COMMISSION, http://reboot.fcc.gov/spectrumdashboard/searchSpectrum.seam. While the FCC is charged with managing most of the wireless spectrum, the National Telecommunications and Information Agency (NTIA) manages government-allocated spectrum. *See About NTIA*, NAT'L TELECOMM. & INFO. ADMIN., http://www.ntia.doc.gov/about (giving an overview of the NTIA's activities).

5. *See* NBC v. United States, 319 U.S. 190, 212–13 (1943) (describing scarcity as a rationale for government control).

6. 47 U.S.C. § 307(c)(1).

7. *Id.* § 307(a)–(b).

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

9. *See* Radio Act of 1927, Pub. L. No. 69-632, § 1, 44 Stat. 1162, 1162 (repealed 1934 and 1966) ("[T]his Act is intended to regulate all forms of interstate and foreign radio transmissions and communications ... and to provide for the use of such channels, but not the ownership thereof...."). The Act made it clear that these licenses should not "be construed to create any right, beyond the terms, conditions, and periods of the license." *Id.*

10. 47 U.S.C. § 309(h).

11. BENJAMIN ET AL., *supra* note 2, at 77–78.

12. 47 U.S.C. § 309(h); *id.* § 310(d).

13. The scarcity rationale, discussed below, has been employed to justify content requirements that likely would be violations of the First Amendment in other contexts. *See* Red Lion Broad. Co. v. FCC, 395 U.S. 367, 390 (1969) (dismissing First Amendment concerns on the basis of spectrum scarcity); *see also* Reed E. Hundt, *The Public's Airwaves: What Does the Public Interest Require of Television Broadcasters?*, 45 DUKE L.J. 1089, 1091–92 (1996) (arguing that children's educational

Pursuant to the Communications Act's "public convenience, interest, or necessity"[14] mandate, broadcast television stations act as "public trustees," whereby broadcasters choose to "sacrifice[] financial gain to serve the interests of the viewing and listening public."[15] For example, over-the-air broadcasters must agree not to transmit obscene material at any hour and indecent material between 6 a.m. and 10 p.m.,[16] television broadcasters must air at least fourteen hours of children's television per week at given times,[17] advertising on children's television must be limited to 10.5 minutes per hour on weekends and 12 minutes per hour on weekdays,[18] and broadcasters must provide equal air time for all legally qualified candidates.[19] As property owned by the people, held in trust by the government, and regulated by the FCC, spectrum continues to be managed in service of the public interest.

While this command-and-control regime has been accepted as the status quo, its justifications have been widely criticized.[20] Even with restrictions, licenses are incredibly valuable.[21] Despite their value, licenses were originally assigned free of charge. When two rivals vied for the same license, the Commission held comparative hearings to determine who would prevail.[22]

---

television, indecency, and political campaign requirements were instituted out of concern that the open market would create a "race to the bottom" vis-à-vis indecency and violence).

14. Communications Act of 1934, Pub. L. No. 73-416, § 303, 48 Stat. 1064, 1082 (codified at 47 U.S.C. § 303).

15. BENJAMIN ET AL., *supra* note 2, at 181; *see Red Lion Broad. Co.*, 395 U.S. at 394 (describing "scarce radio frequencies as proxies for the entire community" and explaining that "the Federal Radio Commission considered the needs of competing communities and the programs offered by competing stations to meet those needs").

16. 47 C.F.R. § 73.3999 (2005); *see also* FCC v. Pacifica Found., 438 U.S. 726, 748–51 (1978) (holding that FCC indecency regulations did not count as censorship under the First Amendment and indecency could be regulated during certain hours).

17. 47 C.F.R. § 73.4050 (1997) (citing FCC, CHILDREN'S TELEVISION REPORT AND POLICY STATEMENT, 50 F.C.C. 2d 1, 1–2 (1974)).

18. 47 U.S.C. § 303a(b).

19. *Id.* § 315(a). However, the equal opportunity doctrine exempts appearances by candidates on bona fide newscasts, interviews, documentaries, and on-the-spot news events. *Id.*

20. *See, e.g.*, Jerry Brito, *The Spectrum Commons in Theory and Practice*, 2007 STAN. TECH. L. REV. 1, ¶ 41 (stating that academics and policy makers view command-and-control regulation as "undeniably inefficient" and seek to identify alternative management systems); Philip J. Weiser & Dale N. Hatfield, *Policing the Spectrum Commons*, 74 FORDHAM L. REV. 663, 668–69 (2005) (describing "the generations old 'command-and-control' model" as "tightly prescrib[ing] what users can and cannot do with a spectrum license" and often preventing "'win-win' trades from taking place").

21. *See* Tom Hazlett, *Putting a Price Tag on TV Spectrum*, TV NEWS CHECK (Nov. 25, 2009), http://www.tvnewscheck.com/article/37599/putting-a-price-tag-on-tv-spectrum (hypothesizing that spectrum allocated for broadcasting is worth $107 billion).

22. *See* 47 U.S.C. § 309(a) ("[T]he Commission shall determine, in the case of each application filed with it . . . , whether the public interest, convenience, and necessity will be served by the granting of such application . . . ."); *id.* § 309(d) ("Any party in interest may file with the Commission a petition to deny any application . . . ."); *id.* § 309(e) ("If . . . a substantial and material question of fact is presented . . . [the Commission] shall formally designate the application for hearing . . . [and a]ny hearing subsequently held upon such application shall be a full hearing in which the applicant and all other parties in interest shall be permitted to participate."); *see also*

By the early 1980s, comparative hearings failed to serve their purpose.[23] Harnessing market forces to properly assign new licenses, Congress first authorized the Commission to assign licenses to previously unassigned spectrum by auction in 1993.[24] By 1997, auctions were mandatory.[25] According to the statute, the goals of these spectral auctions are to promote: "economic opportunity and competition," the ready "accessib[ility]" of "new and innovative technologies," "the development and rapid deployment of new technologies," "efficient and intensive use of the electromagnetic spectrum," and the "recovery for the public of a portion of the value of the public spectrum resource."[26] Whereas licensees initially received valuable broadcasting rights free of charge, competitive bidding has produced tens of billions of dollars for the U.S. Treasury.[27] Recent spectrum auctions brought in about $19 billion for the Treasury.[28] However, auctions have not solved the inefficiencies in the command-and-control regime, partially because they do not allow spectrum already tethered to old technologies to be reallocated.

B. Command-and-Control Concerns

While auctions have been helpful in efficiently allocating unassigned spectrum, the command-and-control regime has left previously assigned spectrum inefficiently allocated. These market inefficiencies, predicted by Ronald Coase in his famous 1959 article critiquing the rationale for government control and calling for private rights in spectrum,[29] do not allow

---

Ashbacker Radio Corp. v. FCC, 326 U.S. 327, 333 (1945) (holding that "where two *bona fide* applications are mutually exclusive the grant of one without a hearing to both deprives the loser of the opportunity which Congress chose to give him").

23. Mark W. Munson, *A Legacy of Lost Opportunity: Designated Entities and the Federal Communications Commission's Broadband PCS Spectrum Auction*, 7 MICH. TELECOMM. & TECH. L. REV. 217, 220–22 (2001) (describing how the failure of comparative hearings compelled Congress to use lotteries to allocate spectrum in 1981 and later revisit spectrum licensing in 1993 to allow competitive bidding).

24. Omnibus Budget Reconciliation Act of 1993, Pub. L. No. 103-66, § 6002, 107 Stat. 312, 387–97 (codified as amended at 47 U.S.C. § 309).

25. 47 U.S.C. § 309(i)–(j).

26. *Id.* § 309(j)(3)(A)–(D).

27. BENJAMIN ET AL., *supra* note 2, at 176 (citing *Auctions Summary*, FED. COMM. COMMISSION, http://wireless.fcc.gov/auctions/default.htm?job=auctions_all (last updated Feb. 1, 2013)).

28. *Wireless Spectrum Auction Raises $19 Billion*, DEALBOOK, N.Y. TIMES (Mar. 19, 2008, 7:54 AM), http://dealbook.nytimes.com/2008/03/19/wireless-spectrum-auction-raises-19-billion/.

29. *See* R.H. Coase, *The Federal Communications Commission*, 2 J.L. & ECON. 1, 14, 29–30 (1959) (arguing spectrum was not uniquely scarce and therefore did not require special government control and advocating for a private rights scheme analogous to real estate, where land is bought and sold privately but still subject to zoning regulation to ensure efficient results). Similar analogies have proposed building on Coase's theory. *See* KRATTENMAKER & POWE, *supra* note 2, at 18 (analogizing to a situation where "paper is scarce" and "[a] Federal Paper Commission would then be necessary to decide how much paper would be available for (say) books and how much for (say) wallpaper . . . [and] who was permitted to engage in book publishing"). Economists generally agree that "Coase's indictment of government spectrum management has largely been vindicated." Brito, *supra* note 20, at ¶ 6.

"[c]ommercially licensed spectrum [to] . . . move efficiently to the use valued most highly by markets and consumers."[30] The power of the market, used to facilitate the flow of almost all other resources to their most highly valued uses,[31] is therefore helpless to reallocate the "scarce" resource to where it is needed most.[32] In a technologically stagnant world, this would not be a problem. However, "rapid technological advances, changing consumer demands, and new market developments steadily erode the utility of spectrum-management decisions that the Commission made years prior to deployment."[33] With authority to repurpose and relicense spectrum at will, the FCC could begin to remedy the misallocation of spectrum. Ironically, the FCC's command-and-control power of assignment does not extend to repurposing, i.e., there is no administrative fiat for relicensing.[34]

The FCC cannot revoke a license just because the spectrum is better suited for another technology.[35] Rather, there must be a "willful or repeated" violation of the license's terms.[36] While this allows the FCC to regulate broadcasting "by raised eyebrow[s],"[37] it is not a useful tool in the fight against spectrum inefficiency. The FCC does have some power to

30. FCC, CONNECTING AMERICA: THE NATIONAL BROADBAND PLAN 79 (2010) [hereinafter NATIONAL BROADBAND PLAN]; *see id.* ("For example, a megahertz-pop may be worth a penny in one industry context and a dollar in another.").

31. *See* KRATTENMAKER & POWE, *supra* note 2, at 18 (explaining that "by adopting public ownership of the spectrum and administrative control over its uses, Congress chose a legal regime for broadcasting that differs radically from the law that governs every other mass communications medium in the United States").

32. *See* SPECTRUM POLICY TASK FORCE, FCC, REPORT OF THE SPECTRUM RIGHTS AND RESPONSIBILITIES WORKING GROUP 6 (2002) (describing the view that "transferability" is "necessary for efficiently allocating any scarce resource among competing uses").

33. *Id.* at 3.

34. There have been proposals to "upgrade" television licenses, essentially giving current licensees the opportunity to use their spectrum for *either* mobile wireless or television broadband. This approach, advocated for by Evan Kwerel and John Williams in a 2002 working paper, would also free up spectrum to be traded on the market to be put at its highest value. *See* Evan Kwerel & John Williams, *A Proposal for a Rapid Transition to Market Allocation of Spectrum* iv (Office of Plans & Policy, FCC, OPP Working Paper No. 38, 2002) (arguing "[m]arkets can move spectrum to its highest value use both now and in the future. . . . [by] restructuring . . . presently assigned and unassigned spectrum into flexible packages of rights that can be readily traded in the marketplace"). However, it is deficient on two accounts. First, while it would please broadcasters who essentially "would receive a free option to use their licenses for more lucrative mobile broadband," holders of mobile broadband spectrum "would object that the dramatic increase in supply of spectrum would decrease the value of their spectrum" that they had paid so much more for since the 1990s. J. Armand Musey, *How the Traditional Property Rights Model Informs the Television Broadcasting Spectrum Rationalization Challenge*, 34 HASTINGS COMM. & ENT. L.J. 145, 165–66 (2012). Second, by upgrading licenses, the federal government would not receive auction proceeds. BENJAMIN ET AL., *supra* note 2, at 88; Musey, *supra*, at 166.

35. *See* 47 U.S.C. § 312(a) (2006) (listing the circumstances under which the FCC may revoke a license).

36. *Id.*

37. BENJAMIN ET AL., *supra* note 2, at 124.

revisit or revise spectrum allocations (not assignments),[38] but this process is frustratingly slow, taking six to thirteen years to clear and reallocate spectrum.[39]

Also, the FCC cannot simply choose not to renew licenses when they expire. The expiration of a license does not trigger a new competitive bidding process for that spectrum.[40] Instead, upon proper application by an incumbent, the FCC must renew the license if it finds "the station has served the public interest, convenience, and necessity" and "there have been no serious violations . . . which, taken together, would constitute a pattern of abuse."[41] The Commission will only deny renewal if "a licensee has failed to meet the requirements . . . and . . . no mitigating factors justify the imposition of lesser sanctions."[42] It is important to note that the Commission may "not consider whether the public interest, convenience, and necessity might be served by the grant of a license to a person other than the renewal applicant."[43] In this way, the Commission may not deny renewal simply because that spectrum has a different, higher value use.[44] Furthermore, licensees have various legal remedies for license denials.[45] In over seventy-five years, the FCC has only denied four renewal applications and has not denied a single one in the last thirty years.[46] The Commission's hands are tied: they cannot move spectrum to new, exciting, and high-value technology.

Furthermore, to deny permits would be bad policy. Incumbent licensees' "renewal expectanc[y]" is partly predicated on the viewing public's reliance on broadcast service.[47] Disruption in quality service harms

38. *See id.* at 88 (discussing the FCC's ability to reallocate spectrum from one specified use to another).

39. NATIONAL BROADBAND PLAN, *supra* note 30, at 79; *see also* Jeffrey A. Eisenach, *Spectrum Reallocation and the National Broadband Plan*, 64 FED. COMM. L.J. 87, 114–16 (2011) (listing examples of administrative delay).

40. 47 U.S.C. § 309(k)(1).

41. *Id.*

42. *Id.* § 309(k)(3).

43. *Id.* § 309(k)(4).

44. STEVEN WALDMAN, FCC, THE INFORMATION NEEDS OF COMMUNITIES 285 (2011) (explaining that the inability of the Commission to compare the public benefit of licensing spectrum to an existing applicant and the public benefit of licensing spectrum to a prospective applicant "*eliminated competition* for licenses").

45. *See, e.g.*, FCC v. Nextwave Pers. Commc'ns Inc., 537 U.S. 293, 295, 304 (2003) (disallowing revocation of the license because doing so would be in violation of the Bankruptcy Code); Trinity Broad. of Fla., Inc. v. FCC, 211 F.3d 618, 631–32 (D.C. Cir. 2000) (refusing to permit cancellation of the license where the FCC's "regulations and other policy statements are unclear, where the [licensee's] interpretation is reasonable, and where the agency itself struggles to provide a definitive reading of the regulatory requirements").

46. WALDMAN, *supra* note 44, at 286–87. *But see* BENJAMIN ET AL., *supra* note 2, at 130 (citing multiple nonrenewals).

47. *See* FCC v. Nat'l Citizens Comm. for Broad., 436 U.S. 775, 805 (1978) (explaining that "preserving continuity of meritorious service furthers the public interest, both in its direct consequence of bringing proved broadcast service to the public, and in its indirect consequence of

viewers. Also, any legitimate risk to nonrenewal would be a disincentive for licensees to invest in quality broadcast equipment and programming.[48] On the basis of these two concerns, licensees can expect the license to be renewed.[49]

Finally, current licensees do not have an incentive to use spectrum efficiently. Television broadcasters were originally given 6 MHz (megahertz) channels on which to broadcast, sufficient to transmit five to six standard definition streams and likely up to two high definition (HD) streams.[50] Many broadcasters still don't use the entire 6 MHz sliver. Of the 294 MHz allocated to television uses across the nation, only 17% is actually being used to broadcast.[51] Without an ability to subdivide and sell the unused spectrum, broadcasters sit on their valuable, unused portions. In this way, much of the spectrum allocated to private commercial use is wasted.

Hampered by command-and-control licensing,[52] the spectrum market is stunted and the economy is harmed. The National Broadband Plan notes that "[s]ome economists estimate that the consumer welfare gains from spectrum may be 10 times the private value to the spectrum holder. If this rule of thumb is true, it suggests that the social value of licensed mobile radio spectrum alone in the United States is at least $1.5 *trillion*."[53] The proper allocation of this resource will help society realize these valuable financial gains. In the past, unleashing new or previously licensed spectrum has also led to unprecedented technological innovation. A prime example of spectrum's value-enhancing capabilities is the Personal Communications Service (PCS) auctions. Spectrum originally allocated to television channels 70–83, auctioned off in the 1990s, was the catalyst for the invention and proliferation of modern mobile cell phone technology.[54] To put it simply,

---

rewarding—and avoiding losses to—licensees who have invested the money and effort necessary to produce quality performance").

48. *Id.*; *see also* Musey, *supra* note 34, at 165 ("[T]he nonrenewal approach .... would disincentivize [spectrum license holders] to bid the highest rates at FCC auctions and invest in the aggressive build out of the very advanced broadband services the FCC seeks to encourage.").

49. *See Nat'l Citizens Comm. for Broad.*, 436 U.S. at 805 (explaining how licensees have "a 'legitimate renewal expectanc[y]'" if they have provided "meritorious service").

50. *See* FCC, *Spectrum Analysis: Options for Broadcast Spectrum* 15–19 (FCC, OBI Technical Paper No. 3, 2010) [hereinafter *Spectrum Analysis*] (explaining "that two stations could voluntarily broadcast HD streams simultaneously over a single six-megahertz channel" and that up to six standard definition stations could share a 6 MHz channel).

51. *See* Thomas W. Hazlett, Unleashing the DTV Band: A Proposal for an Overlay Auction 5–6 (Dec. 18, 2009) (unpublished manuscript), *available at* http://mason.gmu.edu/~thazlett/pubs/NBP_PublicNotice26_DTVBand.pdf (noting that "[t]here are about 1,750 full-power TV stations, yet there are about 10,290 local channel slots," leaving only 17% of television channels used for broadcast).

52. NATIONAL BROADBAND PLAN, *supra* note 30, at 78.

53. *Id.* at 79 (footnote omitted).

54. *See id.* at 78 ("The number of wireless providers increased significantly in most markets. The per-minute price of cell phone service dropped by 50%. The number of mobile subscribers

reallocating broadcast television spectrum to mobile telephony changed the American way of life as we knew it.[55] With a lack of available administrative remedies to repurpose spectrum, the nation is losing out on valuable technology and innovation.

C. *The Spectrum Crunch*

Over wireless networks, smartphones and tablets offer full Internet functionality—allowing users to download data-heavy videos and mobile "apps."[56] These transmissions require more bandwidth (and spectrum) than a conventional telephone conversation, and thus burden networks originally designed solely for voice-to-voice communication.[57] With the increase in wireless data usage, and the resulting strain on spectrum allocated to these uses, industry experts are in agreement that the nation faces a "spectrum crunch."[58]

The recent increase in wireless data usage is staggering. In 2010, the FCC reported that a survey "found that smartphone penetration is now at 33% of mobile subscribers across the four largest wireless operators."[59] A recent article "estimates nearly 116 million Americans will use a smartphone at least monthly by the end of this year, up from 93.1 million in 2011. By 2013, they will represent over half of all mobile phone users, and by 2016, nearly three in five consumers will have a smartphone."[60] Young people are relying on mobile broadband at an even more alarming rate. Pew Research found that 81% of young adults (ages 18–29) use wireless Internet.[61] The data is telling—Americans are becoming more accustomed (and addicted) to the convenience of ubiquitous connectivity, and the trend is likely to continue.

---

more than tripled. Cumulative investment in the industry more than tripled from $19 billion to over $70 billion." (footnotes omitted)); Thomas W. Hazlett, *Hostage Standoff*, AM. ENTERPRISE INST. (Mar. 19, 2001), http://www.aei.org/article/economics/hostage-standoff/ (noting that "channels 70-83 were converted to mobile phone bands").

55. *See generally* JARICE HANSON, 24/7: HOW CELL PHONES AND THE INTERNET CHANGE THE WAY WE LIVE, WORK, AND PLAY (2007) (chronicling the influence of cellular technology on American norms).

56. NATIONAL BROADBAND PLAN, *supra* note 30, at 49, 76–77.

57. *See id.* at 77 ("[S]martphones such as the iPhone can generate 30 times more data traffic than a basic feature phone, and . . . a laptop can generate many times the traffic of a smartphone.").

58. The term spectrum crunch has been adopted by the media to describe the lack of spectrum available to meet the needs of the burgeoning wireless broadband market. *See e.g.*, *The Spectrum Crunch*, CNN MONEY, http://money.cnn.com/technology/spectrum-crunch/ (devoting an entire portion of their technology section to the spectrum crunch).

59. NATIONAL BROADBAND PLAN, *supra* note 30, at 77.

60. *The 'Smartphone Class': Always On, Always Consuming Content*, EMARKETER (May 2, 2012), http://www.emarketer.com/Article/Smartphone-Class-Always-On-Always-Consuming-Content/1009014.

61. AMANDA LENHART ET AL., PEW RESEARCH CTR., SOCIAL MEDIA & MOBILE INTERNET USE AMONG TEENS AND YOUNG ADULTS 4 (2010).

An increase in data traffic accompanies this increase in mobile wireless usage, and it likely will only get worse as Internet apps, social networking sites, and HD video streaming get more complex. Cisco Systems, the Yankee Group, and Coda Research projected that data traffic would be thirty-five times higher in 2014 than in 2009.[62] The FCC reported that "[d]ata traffic on AT&T's mobile network . . . is up 5,000% over the past three years, a compound annual growth rate of 268%."[63] The report continued by noting that in 2009 wireless networks carried an amount of data equivalent to 1,700 Libraries of Congress per month.[64] Furthermore, as the capabilities and speed of mobile devices increase, so does the amount of data used per phone. Between the first quarter of 2009 and the second quarter of 2010, average data usage per line increased almost fivefold.[65] These numbers only relate to smartphone usage. Aircards, devices enabling laptop computers to connect to wireless networks, consume even more data.[66] With more individuals using their phones for Internet, America's reliance on mobile broadband is not waning.

Without sufficient spectrum allocation, wireless companies will not be able to meet consumer demand. The amount of spectrum at any given time is finite.[67] With only a limited amount of spectrum licensed for mobile broadband uses, the increase in data usage is compromising the resource. By 2014, there will likely be a 275 MHz deficit in available spectrum.[68] Not only does this spectrum dearth jeopardize the future of the mobile broadband industry, it currently can lead "to dropped calls, delayed connections, and slower flows of data to mobile devices."[69] To preserve the efficacy of the existing network, companies have begun to put caps on wireless usage or charge high overage rates.[70] With available spectrum to auction to wireless companies, the problem could be easily mitigated—more spectrum could be

62. FCC, MOBILE BROADBAND: THE BENEFITS OF ADDITIONAL SPECTRUM 9 (2010).

63. NATIONAL BROADBAND PLAN, *supra* note 30, at 76 (footnote omitted).

64. *Id.* The report added, "[b]y 2014 . . . North America will carry some 740 petabytes per month, a greater than 40-fold increase." *Id.* at 76–77.

65. COUNCIL OF ECON. ADVISERS, EXEC. OFFICE OF THE PRESIDENT, THE ECONOMIC BENEFITS OF NEW SPECTRUM FOR WIRELESS BROADBAND 3–4 (2012) [hereinafter BENEFITS OF NEW SPECTRUM].

66. NATIONAL BROADBAND PLAN, *supra* note 30, at 77.

67. *See* H.R. REP. NO. 103-111, at 247 (1993) ("[S]pectrum is a non-depletable natural resource and has finite boundaries."); *see also* Ellen P. Goodman, *Spectrum Rights in the Telecosm to Come*, 41 SAN DIEGO L. REV. 269, 285 (2004) (expounding that "[s]pectrum is simultaneously finite and renewable, everlasting and degradable").

68. David Goldman, *Sorry, America: Your Wireless Airwaves Are Full*, CNN MONEY (Feb. 21, 2012), http://money.cnn.com/2012/02/21/technology/spectrum_crunch/index.htm.

69. BENEFITS OF NEW SPECTRUM, *supra* note 65, at 5.

70. Goldman, *supra* note 68; *see also* BENEFITS OF NEW SPECTRUM, *supra* note 65, at 7 (noting that "three of the four largest U.S. wireless carriers have announced that they will be eliminating their unlimited data plans in favor of tiered usage-based pricing").

devoted to wireless broadband to support increased bandwidth needs.[71] However, the FCC has no more suitable unlicensed spectrum to auction.[72]

D. The FCC's Recognition of the Need to Repurpose

The FCC and Congress have taken notice of the nation's spectrum deficit. Through the 2000s, the FCC examined competing spectral demands and in 2010 released the National Broadband Plan (the Plan).[73] The Plan candidly acknowledges, validating the Coasian critique, that "[t]he current spectrum policy framework sometimes impedes the free flow of spectrum to its most highly valued uses."[74] The Plan takes a deregulatory approach, calling on market forces, rather than command and control, to do most of the work to ensure proper allocation in the future.[75] Most famously, the Plan recommended repurposing 500 MHz to broadband use within the next ten years, 300 MHz of which should be made available within five years.[76] The Plan advocates that of the 500 MHz to be reallocated, 120 MHz should come from broadcast television bands.[77]

Key provisions of the Plan call for increased "flexible licensing."[78] A flexible license holder owns an interest similar to that of fee simple in spectrum.[79] As opposed to the current command-and-control licensing scheme that rigidly mandates spectrum uses, holders of flexible licenses are free to use their licensed spectrum for virtually any desired technology and also may transfer their licenses in a secondary market.[80] Exclusive use rights are only limited by the responsibility not to interfere with the rights of other licensees, including limitations imposed to reduce harmful interference.[81] Thus, flexible licenses allow licensees "the maximum possible autonomy to determine the highest valued use of their spectrum."[82] By allowing licensees the freedom to use or transfer spectrum as they see fit, a flexible licensing scheme ensures that spectrum will no longer be tethered to

---

71. More spectrum is not the only tool to combat these problems. Wireless carriers can also increase the efficiency of their wireless technologies and the number of cell cites. Jessica Elder, *Voluntary Incentive Auctions: The Benefits of a Market-Based Spectrum Policy*, 20 COMMLAW CONSPECTUS 163, 171 (2011).

72. Bill Lake, Chief, Media Bureau, Fed. Commc'ns Comm'n, The FCC's Incentive Auction Proposal: New Options for Broadcasters 3–4 (Feb. 28, 2011); Larry Downes, *Averting a Spectrum Disaster: Now for the Hard Part*, CNET (Feb. 25, 2012), http://news.cnet.com/8301-1035_3-57385202-94/averting-a-spectrum-disaster-now-for-the-hard-part/.

73. NATIONAL BROADBAND PLAN, *supra* note 30.

74. *Id.* at 78.

75. *Id.* at 79.

76. *Id.* at 75.

77. *Id.* at 76.

78. *Id.* at 78–79.

79. SPECTRUM POLICY TASK FORCE, *supra* note 32, at 6.

80. *Id.*

81. *Id.*

82. *Id.* at 16.

antiquated uses.[83] However, before flexible licenses can be issued, spectrum must be reallocated.

The broadcast television bands, both UHF and VHF, are most valuable for mobile broadband usage because of their good propagation characteristics[84] and because they are located adjacent to spectrum already allocated to wireless broadband.[85] Some have even dubbed UHF "beach front property" for its ability to easily penetrate buildings, making it best suited to urban areas.[86] The VHF band is most attractive to mobile broadband because, in addition to its propagation characteristics, it is close to the 700 MHz band, which was recently reallocated and auctioned for mobile broadband use after the digital TV (DTV) transition.[87] Also, much of the 6 MHz slices allocated to over-the-air television (OTA TV) broadcasters are inefficiently used or sit idle. One study found that at a given time in New York, broadcasters were only transmitting over 13% of spectrum allocated for their use.[88] Additionally, to be discussed more fully in a later subpart, the economic value of the spectrum used for OTA TV is much lower than the value for mobile broadband.[89] For these and other reasons discussed below, the spectrum currently allocated to OTA TV is particularly attractive for repurposing.

E. A Solution: Voluntary Incentive Auctions

To balance competing needs,[90] the Plan recommends authorizing a novel repurposing mechanism: voluntary incentive auctions.[91] Supported by some of the nation's top economists,[92] voluntary incentive auctions arguably provide the best approach to reallocation because they harness market powers to ensure spectrum is allocated properly; "[p]ut simply, voluntary

---

83. NATIONAL BROADBAND PLAN, *supra* note 30, at 78–79.

84. *Id.* at 88.

85. *Spectrum Dashboard*, FED. COMM. COMMISSION, http://reboot.fcc.gov/spectrumdashboard/searchSpectrum.seam.

86. Nat'l Ass'n of Broadcasters, *Spectrum 101: Are There Different Types of Spectrum?*, FUTURE OF TV, http://www.thefutureoftv.org/spectrum101/differentTypes.asp.

87. Adam LaMore, *The 700 MHz Band: Recent Developments and Future Plans*, DEP'T COMPUTER SCI. & ENGINEERING, WASH. U. ST. LOUIS, http://www.cse.wustl.edu/~jain/cse574-08/ftp/700mhz/index.html (last modified Apr. 21, 2008).

88. *See* Philip J. Weiser, *The Untapped Promise of Wireless Spectrum* 8 (The Hamilton Project, Discussion Paper No. 2008-08, 2008) (citing a study that reported that "during a four-day period in New York City, only 13 percent of spectrum between 30 MHz and 2.9 GHz [where TV is allocated] was occupied at one time or another").

89. *See infra* subpart IV(A).

90. *See* Elder, *supra* note 71, at 173–79 (describing differing views on and against instituting voluntary incentive auctions).

91. NATIONAL BROADBAND PLAN, *supra* note 30, at 90–91.

92. Letter from Paul Milgrom, Gregory Rosston & Andrzej Skrzypacz, Stanford Univ., to President Barack Obama (Apr. 6, 2011), *available at* http://siepr.stanford.edu/system/files/shared/Letter_to_obama.pdf.

incentive auctions assure that spectrum will be reallocated only when its proposed new use is more highly valued than its existing use."[93]

Although the specific design of voluntary incentive auctions is complex, in its basic form a voluntary incentive auction consists of two steps. The first step is known as a reverse auction. The purpose of the reverse auction is to ascertain the amount of money a broadcaster would accept in return for relinquishing its spectrum.[94] This amount is determined through confidential, competitive bidding.[95] In this way, the broadcasters have the power to voluntarily set their own selling price.[96] The FCC would then repack the remaining broadcasters to maximize the continuity of available spectrum.[97] This repacking process would allow the FCC to determine how much spectrum is available and at what cost.[98]

The second step in the process consists of a forward auction. Forward auctions work much the same way as the spectrum auctions regularly run by the Commission. Auction participants, likely mobile wireless companies, would bid for flexible licenses to the spectrum relinquished by the broadcasters.[99] Bids must meet the auction's reserve price, a threshold amount set by the FCC. This price ensures auction proceeds cover the broadcaster's selling price, determined in the reverse auction, while simultaneously leaving sufficient revenue to be shared with the U.S. Treasury.[100]

The benefits of incentive auctions as a reallocation mechanism are numerous. In addition to summoning market forces to reallocate licenses for flexible use, auctions ensure that broadcasters are protected. The purely voluntary nature of the auction means that broadcasters can evaluate the economic benefits of relinquishing their spectrum for themselves.[101] Broadcasters cannot be "evicted." If broadcasters decide their spectrum is

93. Coleman Bazelon et al., An Engineering and Economic Analysis of the Prospects of Reallocating Radio Spectrum from the Broadcast Band Through the Use of Voluntary Incentive Auctions 5 (Sept. 19, 2011), (unpublished manuscript) *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1985691.

94. *Id.* at 8.

95. *Id.* at 28–30. There must be at least two broadcasters bidding. *See id.* at 30 (explaining that there must be more bidders than bids).

96. Lake, *supra* note 72, at 7.

97. *See* NATIONAL BROADBAND PLAN, *supra* note 30, at 89–90 (noting that repacking alone could release up to 36 MHz of additional spectrum from broadcast TV bands).

98. *Legislative Hearing to Address Spectrum and Public Safety Issues Before the Subcomm. on Commc'ns & Tech. of the H. Comm. on Energy & Commerce*, 112th Cong. 23, 26 (2011) (statement of Peter Cramton, Professor of Economics, University of Maryland).

99. *See* Bazelon et al., *supra* note 93, at 5 (describing a two-sided voluntary incentive auction where one side "will bid the amount they need to be compensated to give up their current spectrum licenses" and the "prospective users of reallocated spectrum bid for the new spectrum licenses"); *see also id.* at 5 (explaining that more licensed spectrum should be allocated to mobile broadband services).

100. *See id.* at 10.

101. Lake, *supra* note 72, at 6.

more valuable than the reserve price, then they won't sell—theoretically leaving spectrum in its most highly valued use. By allowing the broadcasters to choose, the FCC eliminates the risk of litigation or other legal challenges to spectrum repurposing, license revocation, or license nonrenewal.[102] An additional benefit is that the FCC can aggregate the relinquished spectrum, repack the TV bands, and auction contiguous wavelengths.[103] Contiguous wavelengths are more valuable than fragmented wavelengths.[104] Furthermore, spectrum can be repurposed much more quickly with an incentive auction.[105] Considering that spectrum reallocation has taken six to thirteen years in the past, many believe the spectrum crisis would already have serious detrimental effects by the time the FCC could act under its current authority.[106]

In addition to meeting the needs of mobile broadband providers and broadcasters, the auctions are particularly attractive because they help reduce the national debt.[107] The FCC hypothesizes that the U.S. Treasury will realize $28 billion from the auctions.[108]

III. Incentive Auction Authorization and the Middle Class Tax Relief and Job Creation Act of 2012

A. The Act

On February 22, 2012, President Obama signed the Middle Class Tax Relief and Job Creation Act of 2012, delegating incentive auction authority to the FCC.[109] The Act entrusts the FCC with broad power to design efficient auctions.[110] With incentive auction authority only recently granted,

---

102. *See* NATIONAL BROADBAND PLAN, *supra* note 30, at 81 ("[S]haring of proceeds creates appropriate incentives for incumbents to cooperate with the FCC in reallocating their licensed spectrum to services that the market values more highly."); Elder, *supra* note 71, at 196 ("[T]he FCC is less likely to be sued due to the voluntary nature of its proposed incentive auctions.").

103. NATIONAL BROADBAND PLAN, *supra* note 30, at 81, 89.

104. Mohammed Alotaibi & Marvin A. Sirbu, Spectrum Aggregation Technology: Benefit-Cost Analysis and Its Impact on Spectrum Value 11 (Sept. 24, 2011) (unpublished manuscript), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1985738.

105. NATIONAL BROADBAND PLAN, *supra* note 30, at 82.

106. *E.g.*, Elder, *supra* note 71, at 191 ("[T]he mobile spectrum crisis will have already arrived if the FCC chose to repurpose spectrum under its current statutory authority.").

107. It may have been another solution to simply modify current broadcasters' licenses for flexible use, which could then be traded to mobile broadband providers. However, this would have removed the government payment for the resource the government had previously given to broadcasters at no cost. The voluntary incentive auction may have been pushed through partly because it is so helpful for solving the current national debt crisis. *See* Middle Class Tax Relief and Job Creation Act of 2012, Pub. L. No. 112-96, § 6402, 126 Stat. 156, 224–25 (to be codified at 47 U.S.C. § 309(j)(8)(G)) (directing the funds garnered from incentive auctions after the fiscal year 2022 to deficit reduction).

108. FCC, FISCAL YEAR 2012 BUDGET ESTIMATES 6 (2011).

109. Middle Class Tax Relief and Job Creation Act § 6402.

110. *Id.*

the auction's complex mechanics have yet to be defined.[111] However, the statutory structure itself provides three ways for broadcasters to participate.

Under the first condition, broadcasters can relinquish rights to transmit on their UHF channels[112] in return for licenses to broadcast over VHF channels.[113] The relocated broadcaster could continue service in a new location and still receive proceeds for relinquishing their rights to the previously allocated spectrum. However, the FCC lacks power to forcibly relocate broadcasters from UHF to VHF.[114]

Under the second condition, broadcasters could choose to relinquish their licenses and move to share a channel with another licensee.[115] The 6 MHz chunk of spectrum allocated per license is sufficient to carry multiple TV stations. In this option, two broadcasters would share one 6 MHz channel to simultaneously broadcast two HD stations without interference and with high picture quality.[116] At lower quality streams (SD), more than two shows could be broadcast simultaneously on the same 6 MHz channel.[117] Currently, individual licensees are already broadcasting two HD streams on one 6 MHz slice.[118] As the Plan notes, "[n]umerous permutations are possible, including dynamic arrangements whereby broadcasters sharing a channel reach agreements to exchange capacity to enable higher or lower transmission bit rates depending on market-driven choices."[119] The rules and regulations governing licensees choosing to share would remain the same: each station will still be licensed and operated separately; be subject to the Commission's obligations, rules, and policies; and be subject to public interest program regulations.[120] Interestingly, the shared 6 MHz channel will not be split up into multiple smaller licensed blocks.[121] Rather, the licensees themselves would cooperate to determine a mutually beneficial division.[122]

Finally, licensees have the option to relinquish all transmission rights "without receiving in return any usage rights with respect to another

---

111. *See id.* (placing only two limitations on the incentive auctions and requiring the FCC to notify Congress of its methodology prior to any such auction).

112. UHF is broadcast on channels 14–51, and VHF is broadcast on channels 2–13. *Antennas and Digital Television*, FED. COMM. COMMISSION (May 10, 2011), http://www.fcc.gov/guides/antennas-and-digital-television.

113. Middle Class Tax Relief and Job Creation Act § 6403(a)(2)(B).

114. *Id.* § 6403(b)(3)(A).

115. *Id.* § 6403(a)(2)(C).

116. NATIONAL BROADBAND PLAN, *supra* note 30, at 90.

117. *Id.*

118. *Id.*

119. *Id.*

120. Innovation in the Broadcast Television Bands: Allocations, Channel Sharing and Improvements to VHF, 77 Fed. Reg. 30,423, 30,424 (May 23, 2012) (to be codified at 47 C.F.R. pts. 73 & 76).

121. *Id.*

122. *Id.*

channel."[123] In this option, broadcasters would relinquish their full 6 MHz—which may mean they are awarded a higher reserve price.[124]

In both the first and second options, the benefits for broadcasters are clear. First and foremost, the broadcasters remain on the air. With the infusion of capital resulting from the incentive auction, broadcasters, in their new locations and arrangements, could reinvest in high-quality programming and equipment.[125]

Furthermore, the statute builds in protections for broadcasters choosing the first and second options. First, the FCC will reimburse broadcasters for costs associated with channel relocation.[126] Second, because there will be no new infrastructure changes, broadcasters will not need to install new equipment, nor will consumers be required to purchase new receivers.[127] Third, licensees relocating to a shared channel will retain their must-carry rights.[128] These three protections essentially keep the broadcasters similarly situated, but with capital gains as a reward for cooperating. There will be minimal consumer disruption, minimal broadcaster disruption, and an influx of capital to reinvest.

The Act also ensures the Treasury will realize gains. Flexible licenses will only be issued if two conditions are met. First, the amount realized in the forward auction must compensate the relinquishing broadcasters at the price set in the reverse auction.[129] Second, if the total amount of the proceeds from the forward auction are not greater than the sum of (1) the total amount of compensation to be awarded to successful reverse auction bidders, (2) the costs of conducting the forward auction, including repacking costs, and (3) the relocation costs associated with relocated broadcasters, then no spectrum will be relinquished or reallocated, and the FCC will assign no new flexible licenses.[130] Besides the $1.75 billion set aside to pay for relocation costs and the amounts used to pay the broadcaster's reserve price, auction proceeds will be deposited in the Treasury.[131]

---

123. Middle Class Tax Relief and Job Creation Act of 2012, Pub. L. No. 112-96, § 6403(a)(2)(A), 126 Stat. 156, 225 (to be codified at 47 U.S.C. § 1452).

124. *Id.*

125. *Cf. In re* Innovation in the Broadcast Television Bands: Allocations, Channel Sharing and Improvements to VHF, Notice of Proposed Rulemaking, 25 FCC Rcd. 16,498, 16,505 (Nov. 30, 2010) (explaining that television stations sharing channels can use additional income from these arrangements to enhance their programming).

126. Middle Class Tax Relief and Job Creation Act § 6403(b)(4).

127. Lake, *supra* note 72, at 8.

128. Middle Class Tax Relief and Job Creation Act § 6403(a)(4).

129. *Id.* § 6403(c)(1)(B).

130. *Id.* § 6403(c)(2)(A)–(C).

131. *Id.* § 6402.

B. *Full Relinquishment*

Voluntary incentive auctions provide an optimal market-based mechanism whereby broadcasters can voluntarily submit spectrum to be flexibly repurposed. While the Act will help stave off the spectrum crisis, its structure is not optimal. The first two options—relocation and channel sharing—are accompanied by incentives and protections that make them attractive options. However, the third option—full relinquishment—stands alone. The statute does not include any incentives or protections for broadcasters choosing this option. Granted, an inherent incentive for full relinquishers may be the reserve price compensation itself, which could be higher for full relinquishers who are giving up more spectrum than sharers or relocaters. However, this extra capital may not be enough to coax broadcasters off the air, especially for larger, urban broadcasters with fully entrenched businesses, large investments, and a devoted public following. To dislodge this cross section of licensees, more incentives and protections may be needed.

This rest of this Part will outline the possible reasons why Congress did not provide for incentives or protections for full relinquishers. Part IV will then discuss why full relinquishment is the optimal broadcaster choice.

C. *Disincentives to Relinquish*

While there is little legislative history about incentive auction design, there were likely two reasons analogous protections and incentives were not offered to full relinquishers. First, Congress likely assumed that full relinquishers would not need them. The purpose of the protections afforded relocating and sharing broadcasters is to keep them as similarly situated after the auctions as they were before the auctions.[132] By protecting must-carry rights, for example, the sharing broadcasters would not risk losing local cable audiences on which they rely. Broadcasters choosing to fully relinquish their spectrum, and perhaps leave the air entirely, would not need must-carry protections. Second, broadcasters relinquishing their spectrum, and presumably not relocating to another wavelength, would not need their relocation costs covered. Thus, it is likely that Congress did not perceive a need to provide analogous protections for fully relinquishing broadcasters.

Congress's assumption is faulty. Broadcasters choosing to relinquish their spectrum may actually have opportunities to continue broadcasting. For example, broadcasters "could co-broadcast with another broadcaster, obtain a license to broadcast on VHF channels, modify their coverage area or negotiate to have their programming carried on non-broadcast delivery

132. *See* Lake, *supra* note 72, at 8 (assuring broadcasters that the FCC is "committed to working with [them] to ensure that [their] realignment will be as painless as possible" and asserting that, after the auctions, broadcasters will not have to install new infrastructure).

systems . . . [like] cable, satellite and internet video services such as Hulu."[133] Broadcasters could also choose to fully relinquish and subsequently lease spectrum from other licensees.[134] Also, about 60% of broadcasters currently negotiate to be carried on cable. These broadcasters could choose to relinquish and still maintain their viewing audiences via cable, which accounts for 90% of their viewership.[135] The choice to fully relinquish and renegotiate transmission on a different platform may, for many, make financial sense. The numerous options broadcasters have to fully relinquish and then continue broadcasting make it peculiar that Congress did not include some type of incentive or protection for broadcasters choosing to fully relinquish.

As a result of Congress's faulty assumption, the Act does not incentivize or protect broadcasters choosing to fully relinquish. Without incentives or protections, broadcasters are less likely to fully relinquish their spectrum, and therefore the current statutory scheme is detrimental to the overall goal of reallocating spectrum to its most highly valued use.

IV. Full Relinquishment Is the Optimal Broadcast Choice

The option allowing broadcasters to fully relinquish their licensed spectrum is optimal from a policy perspective. At first glance it seems obvious that full relinquishment makes most sense—more spectrum relinquished means more flexibly licensed spectrum put to high-value use. However, the benefit from releasing additional spectrum from its TV uses must be balanced against what is lost—whereas in the channel-sharing or relocation options viewers will be able to simply change to a different OTA channel, full relinquishment risks program cessation and the resulting negative impact on OTA TV viewers that rely on free access to TV.

A. Economic Benefits of Full Spectrum Relinquishment

First, the full-relinquishment option is optimal for the same reasons incentive auctions themselves are desirable—spectrum tethered to TV broadcasts is not being put to its most highly valued uses. Full relinquishment, therefore, moves the most spectrum (versus the sharing and relocation options) from low to high value.

There is general agreement that the vitality of the broadcast television industry is declining. In 2010, only 10% of total television viewers relied solely on OTA TV, down from 24% in 1999.[136] Prime time network ratings have also declined 25%–30%.[137] Viewers are replacing OTA TV with

133. Bazelon et al., *supra* note 93, at 1–2.

134. *See* Eisenach, *supra* note 39, at 95 (explaining that, since October 2003, the FCC has allowed "the leas[ing] of spectrum usage rights").

135. *Spectrum Analysis*, *supra* note 50, at 7–8.

136. *Id.* at 7.

137. *Id.* at 8.

cable, satellite, and Internet outlets.[138] Despite must-carry privileges, which have arguably kept broadcasters in business,[139] "broadcast TV station revenues have declined 26%, and overall industry employment has also declined."[140] The depressed TV outlook devalues currently licensed spectrum.[141]

While broadcast television is declining, mobile broadband is thriving. Despite the economic downturn, "the mobile wireless industry has been a source of stability and revenue, contributing investment, jobs, and increased productivity to the U.S. economy."[142] Wireless revenues have increased 39%, and employment in the industry has grown 16% between 2005 and 2010.[143]

The boom from general (non-mobile) broadband use[144] has been linked to numerous tangible and intangible economic benefits. The statistical data is extensive,[145] but two studies are of note. First, a 2011 study found that introducing broadband causes gross domestic product (GDP) to increase 2.7%–3.9% per capita.[146] Furthermore, the study found that by increasing broadband penetration by 10% the annual growth rate of per capita GDP increased by 0.9–1.5 percentage points.[147] Second, a 2011 study in Germany found that innovation increases by approximately forty percentage points when access to broadband increases.[148] Spectrum reallocation will likely make mobile broadband less expensive and therefore increase the access to broadband use in mobile form, giving many access to broadband service they would not otherwise have.

The economic impact of the growth in the wireless industry is also beneficial. One study reports that the likely investment of $25–$53 billion

138. *See id.* at 7–8 (discussing the decline of OTA TV and the demand for newer technologies); *see also Broadcasters Worry About 'Zero TV' Homes*, NPR, Apr. 17, 2013, http://www.npr.org/templates/story/story.php?storyId=176496138 (noting that the number of homes without cable, satellite, or traditional OTA TV has increased from 2 million in 2007 to 5 million in 2013 and that this trend negatively impacts broadcaster revenues).

139. *See infra* notes 190–94 and accompanying text.

140. *Spectrum Analysis*, *supra* note 50, at 8 (footnotes omitted).

141. *Id.* at 7.

142. Elder, *supra* note 71, at 170.

143. *Spectrum Analysis*, *supra* note 50, at 7.

144. *See supra* subpart II(C).

145. *See* BENEFITS OF NEW SPECTRUM, *supra* note 65, at 15 (referencing "[a] number of economic studies" reviewing the impact of broadband).

146. Nina Czernich et al., *Broadband Infrastructure and Economic Growth*, 121 ECON. J. 505, 507 (2011).

147. *Id.*

148. *See* Irene Bertschek et al., *More Bits—More Bucks? Measuring the Impact of Broadband Internet on Firm Performance* 12 (Ctr. for European Econ. Research, Discussion Paper No. 11-032, 2011) (reporting that, according to one of its models, the use of broadband Internet increases the likelihood of an innovation by about 40%).

over the next four years could account for $73–$151 billion in GDP growth and up to 771,000 new jobs.[149]

Market valuation metrics reflect the data presented above. It is clear that the promising mobile broadband industry far outweighs the TV economy in terms of economic value. The difference in valuation is staggering. The value of spectrum is measured in terms of dollars per MHz per person reached (dollars per MHz-Pop). A recent study estimates that spectrum allocated for flexible licenses is worth $1.35 per MHz-Pop.[150] The value of spectrum allocated for TV usage is estimated to be up to $0.15 per MHz-Pop.[151] Spectrum is undervalued as currently allocated to TV usages.

Furthermore, increased investment in the mobile broadband industry will lead to far-reaching nonmeasurable innovations and impacts on the economy. One article has explained that mobile broadband investment effects are "similar to building a roadway, which not only generates jobs and income for the builders of the road, but also provides opportunities for others to create new businesses and homes along the roadway."[152] Some intangible, social benefits of mobile broadband spectrum allocation have already been realized. Mobile broadband has brought "many of the breakthroughs that the Internet has fostered in civic engagement and First Amendment expression to new devices . . . and underrepresented populations," for example "innovat[ive] . . . journalism."[153] The nation has also already seen increases in safety and security through location and recovery services.[154] The Executive Office of the President's Council of Economic Advisers predicts other opportunities could include consumer applications (e.g., Apple iPhone apps), increased business productivity (e.g., cloud computing to decrease fixed costs), positive impacts on patient care and decreasing the pace of health care cost growth (e.g., videoconferencing for patients in difficult-to-access areas or living far from specialists), and education (e.g., educational apps and connectivity between students and classrooms).[155] Increasing spectrum for mobile broadband, and thus the opportunity for investment, will likely have a net positive impact on the economy and spur innovation yet to be seen.

As outlined in subpart II(B), the social opportunity cost for the misallocation of spectrum is $1.5 trillion. While this value gap provides a

149. DELOITTE, THE IMPACT OF 4G TECHNOLOGY ON COMMERCIAL INTERACTIONS, ECONOMIC GROWTH, AND U.S. COMPETITIVENESS 7 (2011).

150. Bazelon et al., *supra* note 93, at 23.

151. *Spectrum Analysis*, *supra* note 50, at 7.

152. Elder, *supra* note 71, at 170 (citing Alan Pearce & Michael S. Pagano, *Accelerated Wireless Broadband Infrastructure Deployment: The Impact on GDP and Employment*, 18 MEDIA L. & POL'Y 105, 107 (2009)).

153. *Spectrum Analysis*, *supra* note 50, at 10.

154. *See id.* (noting that "mobile broadband applications can leverage location-based services to improve public safety through faster location and recovery of missing persons and stolen property").

155. BENEFITS OF NEW SPECTRUM, *supra* note 65, at 7–12.

strong argument for reallocation generally, it also provides a strong reason why the full relinquishment option is most attractive. Full relinquishment unleashes as much spectrum as possible, ensuring the largest economic and social gains.

B. Constituency Benefits

The full relinquishment option most benefits the FCC, mobile wireless carriers, and the government. The increased spectrum available as a result of broadcasters choosing the full-relinquishment option, versus sharing or relocation, benefits mobile wireless carriers and the FCC for two reasons. First, and most obviously, it unleashes more spectrum on the market to be allocated for wireless use, leading to more investment in technologies. While analysts are unsure of the exact amount of spectrum necessary for mobile broadband,[156] the opportunity to release tethered spectrum for flexible licenses means that any excess spectrum not used for mobile broadband will still be put to its highest value use. If insufficient spectrum is reallocated in the first round of auctions, perhaps because too many broadcasters decide to share or relocate, further measures would have to be taken in the future—measures that may be less palatable to broadcasters. Furthermore, it is in the FCC's best interest to conduct the smallest number of auctions possible as they are expensive and time-consuming. Therefore, instead of possibly delaying the problem, it would be wise to relinquish as much spectrum as possible now.

Second, full relinquishment will allow more contiguous spectrum to be auctioned. Contiguous spectrum is more valuable to mobile broadband users.[157] If more spectrum is relinquished at one time, then the FCC can repack more efficiently and auction larger, contiguous swaths of spectrum. With fewer broadcasters choosing the full-relinquishment option, there will be less contiguous spectrum auctioned at this time, or if there is a need to conduct these auctions again, more fragments to be auctioned in the future. In both of these scenarios, value is diminished.

Full relinquishment also benefits the government. Assuming the full 120 MHz is auctioned, at $1.35 per MHz-Pop, it is estimated that the auctions could bring in approximately $40 billion in gross revenue for the U.S. Treasury.[158] Using these numbers, it follows that for each additional MHz auctioned the Treasury will realize an extra $333.33 million in revenue. Full relinquishment, releasing the most MHz, will lead to the most

156. *See* Morgan Reed, *Why the Vaunted Spectrum Auctions Won't Cut It*, CNET (Feb. 28, 2012), http://news.cnet.com/8301-1035_3-57386922-94/why-the-vaunted-spectrum-auctions-wont-cut-it/ (arguing that auctions will not provide sufficient spectrum to support mobile broadband growth).

157. *See supra* note 104 and accompanying text.

158. Bazelon et al., *supra* note 93, at 23–24.

government revenue of the three options. Considering the delicate economy and high deficit, increased Treasury funds are crucial.[159]

C. *The Costs of Full Relinquishment*

The productive benefits of full relinquishment, however, must be balanced against the potential drawbacks. While leaving the air is not a foregone conclusion, it is likely that some licensees choosing the full-relinquishment option will cease production. Thus, there will be some harm to viewers relying on free OTA TV.

The FCC has been clear in acknowledging its wish to maintain a strong OTA TV market. The Plan explicitly states, "[b]ecause of the continued importance of over-the-air television, the recommendations in the plan seek to preserve it as a healthy, viable medium going forward, in a way that would not harm consumers overall."[160] Broadcasting over airwaves held in trust for public benefit still provides an important service.

OTA TV provides a free service for American viewers, many of whom "tend to have lower incomes, are more likely to be over age 65, and to live in rural areas."[161] Currently, an estimated 17 million households (46 million people) still access OTA TV.[162] The OTA TV service still provides valuable programming to American audiences. For example, the FCC conditions licensing on the provision of children's programming, reasonable access for federal political candidates, and content restriction, among other things.[163] Local news coverage and emergency notifications, commonly broadcast OTA by local stations, have been especially valuable services, especially at times that other media forms are insufficient.[164]

Most significantly, the increased risk of program cessation in local markets threatens access to important broadcast services. In sharing and relocation conditions, broadcasters would likely continue to broadcast in their local area. In those conditions, where a station was relocated out of an area, a new program would likely take its place. Consumers would not significantly lose out on the free public services they enjoy.[165] In the full-

159. *See U.S. Budget Deficit*, BROOKINGS, http://www.brookings.edu/research/topics/budget-deficit (stating that "[t]he U.S. federal budget deficit continues to grow"); Ron Haskins, *Going Big on Deficit Reduction Is Dead. Now What?*, BROOKINGS (Jan. 14, 2013), http://www.brookings.edu/research/opinions/2013/01/15-small-deficit-deal-haskins (describing the "catastrophic effect" the U.S. budget deficit will have on the U.S. economy).

160. NATIONAL BROADBAND PLAN, *supra* note 30, at 89.

161. Musey, *supra* note 34, at 180–81 (footnotes omitted).

162. Barbara Cochran, *Should Some of Broadcasters' Spectrum Be Auctioned Off to Wireless Carriers? No: It Hurts Local TV*, WALL ST. J. (Nov. 14, 2011, 3:56 PM), http://online.wsj.com/article/SB10001424052970203716204577017801681007194.html.

163. *See supra* notes 13–19 and accompanying text.

164. Cochran, *supra* note 162.

165. *See* NATIONAL BROADBAND PLAN, *supra* note 30, at 90 (observing that some OTA TV consumers "might gain reception from one or more stations as a result of changes to service areas"

relinquishment condition, however, there is a higher risk of significant service loss. Viewers would stand to lose some of the indecency and obscenity requirements, educational programming, and other broadcasting conditions because the FCC could no longer impose the conditions on speech that could be attached to licensure.[166] While cable or satellite could fill some voids, OTA TV is uniquely situated to reflect local values and address local concerns. Although hard to value economically, these local stations add to the character of the community, bringing cities and locales together. Also, in times of emergency with clogged cell phone networks, local broadcast television provides a stream of information and comfort to viewers who would otherwise be uninformed.[167]

Other intangible costs may also be incurred. For example, children without cable or Internet could be ostracized at school for not being kept abreast of popular programming. Adults will be less in tune with local events and, without other outlets, unable to easily find and connect with like-minded community members. Although seemingly petty, these effects are real.

Full relinquishment may have some economic costs as well. Currently, local television employs 1.54 million people.[168] New mobile DTV technology also has the potential to provide valuable service, but it is still in its infancy and it is difficult to project its value.[169] While most broadcasters will not actually leave the air,[170] the full-relinquishment condition could have a significant negative impact on the extent of and access to local broadcast coverage for millions of viewers.

D. On Balance, Full Relinquishment Is Optimal Despite Costs

While the full-relinquishment condition potentially harms viewers, the injury will be minimal and can easily and inexpensively be mitigated. Harm to viewers relying on local OTA TV will be small, only seriously impacting rural viewers who cannot access or afford cable or satellite. First, the number of television viewers relying on OTA TV is already miniscule and declining further. As mentioned previously, only 10% of current viewers solely watch OTA TV, and that number is declining.[171] Between 1999 and

---

and that generally "[c]onsumers would continue to receive over-the-air television" after reallocations of spectrum).

166. *Cf.* United States v. Playboy Entm't Grp., Inc., 529 U.S. 803, 815 (2000) (observing, in the context of decency requirements, that the First Amendment does not allow the same level of regulation of cable television because of its technological differences from broadcast).

167. Cochran, *supra* note 162.

168. *Id.*

169. *See Spectrum Analysis*, *supra* note 50, at 32 (describing the current state of the mobile DTV business model as "nascent").

170. Cochran, *supra* note 162.

171. *See supra* note 136 and accompanying text.

2010, there was a fourteen-percentage point drop in OTA viewership.[172] With the addition of new entertainment and information outlets, this number will likely continue to drop.[173]

Second, the FCC projects that the consumer impact would be most prevalent in urban markets.[174] In large urban areas there is more demand for spectrum because more stations are clogging the airwaves. In smaller markets, where only about 20% of the available channels are occupied,[175] there is less need to vacate the spectrum for mobile wireless use. Therefore, broadcasters in local areas are less likely to be included in an incentive auction by the FCC.[176] One consumer protection envisioned in the Plan is to *not* accept spectrum allocated to important television coverage in rural areas and smaller markets.[177]

Third, while most consumer impact will take place in large, urban areas, these areas are in least jeopardy of realizing the harmful service losses. In these markets, "[c]onsumers . . . tend to have a relatively large number of alternatives to view television content—a median of 16 over-the-air full-power television stations, over-the-air low-power stations and digital multicast channels, at least three to four multichannel video programming distributors (MVPDs), and a growing amount of broadband Internet video content."[178] These markets also tend to have the lowest over-the-air viewership.[179] Thus, the harm to consumers in urban areas, where most broadcasters may leave the air, is minimal.

Fourth, many of the OTA TV stations will have the power to negotiate retransmission contracts with cable providers, so leaving the air will not harm many viewers. In 1992, Congress passed the Cable Television Consumer Protection and Competition Act, requiring cable companies to

---

172. *Spectrum Analysis*, *supra* note 50, at 7.

173. *See id.* at 7–8 (discussing new outlets that are contributing to the decline in OTA TV viewership and the "challenging long-term trends" facing OTA TV in general).

174. *See* NATIONAL BROADBAND PLAN, *supra* note 30, at 90 (explaining that reallocation would most likely occur "in the country's largest, most densely populated markets").

175. *See Spectrum Analysis*, *supra* note 50, at 29 (reporting that 93% of smaller markets "have fewer than 10 channels directly allotted to full-power TV broadcasters (of the 49 channels in total)").

176. *Id.*

177. *Cf.* NATIONAL BROADBAND PLAN, *supra* note 30, at 90 (noting that "the FCC should ensure that consumers in rural areas and smaller markets retain service and are not significantly impacted" and suggesting that the reallocation mechanisms will primarily involve "the country's largest, most densely populated markets, where the greatest demand for spectrum and the greatest congestion within the broadcast TV bands coincide").

178. *Id.* at 90–91.

179. *Cf.* Congresswoman Diane E. Watson, Keynote Address to Minority Media and Telecommunications Council Regulatory Breakfast on Minority Media Ownership and Telecommunications Legislation (July 19, 2005) (explaining that "over-the-air-only households . . . disproportionately include . . . rural households").

retransmit local broadcast stations.[180] The purpose of this legislation was to protect the vitality of local broadcasters falling victim to increased competition from cable and thus losing advertising revenue.[181] Recently, however, many broadcast stations have foregone their must-carry rights and negotiated retransmission contracts in which their content was carried on cable in return for subscriber fees.[182] As of 2009, only 37% of stations still relied on must-carry rights for retransmission.[183] Thus, the majority of stations have the negotiation power to move their programming to cable. While this move will affect OTA TV viewers who cannot afford or do not have access to cable, it will enable broadcasters to continue to serve the 90% of viewers watching local, community content on cable or satellite,[184] thus minimizing the effect of full relinquishment.

Finally, local news and community programming of some sort will still be available over the airwaves. It is likely that broadcasters with the most robust business models and largest audiences, and broadcasters most entrenched in the community, will not choose to relinquish, even with incentives. The long-term revenue from broadcasting will likely outweigh the marginal value of increased spectrum sales. Thus, these broadcasters will likely choose to channel share or relocate, if choosing to participate in auctions at all. Also, increased spectrum allocation and further innovation will lower costs for access to mobile wireless and streaming video, thus replacing many lost OTA TV programs or services.[185]

Viewers in rural areas, or viewers with limited access to cable or satellite, may be significantly harmed by broadcasters choosing full relinquishment for the reasons stated above. However, this harm can be mitigated. For example, the government could provide "lifeline" cable or satellite subscription, whereby harmed consumers would receive all the OTA TV signals in their market and, perhaps, also wired broadband service.[186] The government could also provide coupons for equipment upgrades that could help consumers gain access to distant signals.[187] While these subsidies would help most viewers, many rural communities do not have cable or satellite connections. One commentator suggested that the cost of

180. Cable Television Consumer Protection and Competition Act of 1992, Pub. L. No. 102-385, § 4, 106 Stat. 1460, 1471–77 (codified at 47 U.S.C. § 534 (2006)).

181. *Spectrum Analysis*, *supra* note 50, at 8.

182. *Id.*

183. *Id.*

184. *Id.* at 7.

185. *See* NATIONAL BROADBAND PLAN, *supra* note 30, at 90–91 (discussing alternatives that have the potential to replace OTA TV, like "broadcasting popular video content to mobile devices"). For example, there is extensive local coverage, political commentary, educational applications, and copious amounts of entertainment online. *See id.* at 5 (discussing the importance of broadband access to opportunity and citizenship since so many educational and political discussions occur online).

186. *Id.* at 91.

187. *Spectrum Analysis*, *supra* note 50, at 30.

outfitting rural areas with cable and providing a "local stations only" cable package would be significantly smaller than the government auction proceeds.[188] This package would allow for all Americans to retain access to local television, despite the absence of the full relinquisher's signals. While these services would increase access, they would not replace the content lost by broadcasters who fully relinquish and chose not to retransmit their programming on cable or satellite.

The myriad benefits from the full relinquishment condition outweigh the minimal harm to consumers. Because most broadcasters can continue to transmit programming via satellite or cable, and OTA TV viewers can still retain access to one of these two services, the only real harm will be to viewers who lose the content on their favorite television station. The value of the relinquished spectrum as an economic and technological driver dwarfs the negative impacts to these consumers—impacts that are minimal and can be mitigated.

However, broadcasters may be unwilling to participate in the shared or relocating conditions, let alone choose to fully relinquish.[189] While they may be mistaken that the choice to fully relinquish necessarily harms local communities or that they will be forced to stop programming, this very likely may be the sentiment. Thus, there may need to be more incentives for broadcasters to choose to fully relinquish. To make full relinquishment more attractive to broadcasters, the FCC must create ways to keep broadcasters on the air or to more fully compensate them for discontinuing their broadcasts.

V. Full-Relinquishment Incentives

The benefits of full relinquishment far outweigh the costs, especially considering that many broadcasters will not stop programming once they leave the air. To optimize the potential gains from the voluntary incentive auction, the FCC should incentivize broadcasters to choose to fully relinquish their 6 MHz sliver of spectrum. This Part will outline the potential incentives in the FCC's toolbox.

Increased Auction Proceeds. Naturally, a prime incentive for full relinquishment is a higher auction reserve price. However, this incentive may not be enough for many successful broadcasters who turn a large profit. To incentivize broadcasters who on the margin need some additional financial compensation to relinquish, the FCC could put a premium on this choice. While economists would have to determine the proper formula for this amount, it likely would be a strong incentive.

---

188. Musey, *supra* note 34, at 160–61.

189. Joe Flint, *FCC Can Auction Spectrum, but Will Broadcasters Sell?*, L.A. TIMES (Feb. 17, 2012, 4:33 PM), http://latimesblogs.latimes.com/entertainmentnewsbuzz/2012/02/broadcast-spectrum.html.

Carriage Rights. Much of the OTA TV broadcasters' value lies not in their OTA broadcasting, but in their retransmission via cable or satellite on which most (90%) of their audience watches.[190] As briefly mentioned earlier, federal law requires cable and satellite companies to rebroadcast local OTA content.[191] These rules were designed to protect broadcasters from being driven out of the market by cable companies refusing to show their programming.[192] By preserving the efficacy of local programming, must-carry rules accomplished three goals: saving free OTA TV, promoting a diversity of viewpoints, and ensuring fair competition in the broadcast television market.[193]

While most broadcasters can currently successfully negotiate to be carried on cable,[194] the must-carry rules protect less popular, niche, local programming. Ultimately, these rules protect both the broadcasters—who remain in a competitive position vis-à-vis more popular local broadcasters—and the consumers—who do not lose out on unpopular but meaningful local coverage. Must-carry rights do not apply to nonbroadcast programs—no broadcast license means no must-carry rights.[195] Thus, if broadcasters choose to fully relinquish their spectrum (and thus their license), they will also lose their must-carry privileges.

If broadcasters who are unable to negotiate with cable for transmission could maintain their must-carry privileges, they could continue to broadcast without losing a significant portion of their audience, and communities would not lose the local programming, thus mitigating any negative effects for either entity. Preservation of must-carry rights would be a strong incentive for these broadcasters to choose the full-relinquishment condition. While arguments can be made that fully relinquishing broadcasters can and should keep their must-carry privileges, Congress or the Supreme Court may not be convinced.

In the face of a First Amendment challenge, in *Turner I*[196] and *Turner II*,[197] the Supreme Court upheld the content-neutral must-carry provisions because they furthered an important government interest: survival of the free OTA broadcast medium.[198] A must-carry incentive for full relinquishers would likely also have to withstand a First Amendment challenge. To do so, fully relinquishing broadcasters would have to demonstrate the

190. Musey, *supra* note 34, at 146.

191. 47 C.F.R. § 76.56 (2012). 47 C.F.R. § 76.56 outlines cable must-carry rules. 47 U.S.C. § 338 (2006) provides the analogous satellite retransmission rules.

192. BENJAMIN ET AL., *supra* note 2, at 498.

193. Turner Broad. Sys., Inc. v. FCC (*Turner II*), 520 U.S. 180, 189 (1997).

194. *Spectrum Analysis*, *supra* note 50, at 8.

195. BENJAMIN ET AL., *supra* note 2, at 497–98.

196. Turner Broad. Sys., Inc. v. FCC (*Turner I*), 512 U.S. 622 (1994).

197. *Turner II*, 520 U.S. 180 (1997).

198. *Id.* at 185, 213; *see also Turner I*, 512 U.S. at 661–62 (finding that the must-carry provisions are content neutral).

absence of any content-based distinctions and that their continued operation on cable systems constitutes an important government interest despite not being offered for free OTA.[199] J. Armand Musey provides two rationales under which this could be accomplished. First, he suggests that the full-relinquisher status itself could justify a new, non-content-based distinction that would allow for the must-carry privilege extension:

> Continuation of must-carry requirements after the broadcasters no longer broadcast over-the-air would be fully compatible with *Turner I* and *Turner II* so long as Congress finds another equally valid non-content based distinction to separate the broadcasters from others. One such option would be the creation of "broadcasting licenses" given to former broadcasters if they meet designated content-neutral requirement(s). However, there is a risk the Court could find that such licenses are not valid because they privilege the now former broadcasters' content because of who they are (former broadcasters) as opposed to "the manner in which the speakers transmit their messages to viewers."[200]

His concern may be without merit. In order to qualify for less strict "important government interest" First Amendment scrutiny, broadcasters would have to show they were not being favored on the basis of their content.[201] While Musey understands the distinction he presents to be a problem, under the language of *Turner I*, it actually may be permissible. Whereas in *Turner I* the speakers were distinguished (favored) "based only upon the manner in which speakers transmit their messages to viewers, and not upon the messages they carry,"[202] the new special "broadcast licensees" would be favored on the basis of their choice of full relinquishment in the incentive auction (or as Musey puts it, "who they are"), but *not* their content. Whether a full relinquisher was "commercial or noncommercial, independent or network affiliated, English or Spanish language, religious or secular"[203] would have no bearing on the must-carry privileges. Whether they previously had a license and chose to fully relinquish would be the only basis for awarding must-carry privileges. As such, a choice to extend must-carry privileges would not be pretext for content-based privileges. Where this analysis runs into trouble is the second hurdle—what important government interest the non-content-based distinction would serve.

The important government interest on which the law was upheld was "to guarantee the survival of a medium that has become a vital part of the Nation's communication system, and to ensure that every individual with a

199. Musey, *supra* note 34, at 155.
200. *Id.* at 156–57 (footnote omitted).
201. *Turner I*, 512 U.S. at 642–43.
202. *Id.* at 645.
203. *Id.*

television set can obtain access to free television programming."[204] The full-relinquishment condition clearly does not serve this important governmental interest—in the full-relinquishment condition, broadcasters are being explicitly removed from the airwaves. Thus, another important government interest must be served.

Musey argues that a potential important government interest, articulated in Justice Breyer's *Turner II* concurrence, is "to assure the over-the-air public 'access to a multiplicity of information sources.'" [205] Musey contends:

> The important government interest would be the promotion of widespread access to local television content for cable and satellite subscribers as well as over-the-air viewers. Local television broadcasters are a primary source of local news content for many people, regardless of how they receive their television signals. The benefits of diversity of content, particularly local content, could not be fully maintained without keeping the current broadcasters in business via transmission on cable and satellite systems.[206]

The Court would then have to find the law narrowly tailored to reach this government interest.[207] Musey's interpretation could be the rationale necessary to uphold must-carry privileges as applied to fully relinquished broadcasters. Additionally, the significance of flexibly licensing the airwaves for mobile wireless use and the economic benefits that attach could provide another government interest sufficiently important to satisfy the First Amendment test.

Though convincing, both rationales may prove inadequate. The efficacy of the *Turner II* decision itself has been widely criticized, and many believe it could not withstand an attack as applied currently, let alone to nonbroadcasting programmers.[208] When *Turner I* was decided, 40% of Americans lived without cable and relied on OTA TV.[209] Now, with only 10% living without cable,[210] the importance of "preserving access to free television"[211] is less important. As Musey puts it, "[a]s a result [of the diminished reliance on OTA TV], one of the primary justifications for upholding must-carry regulation has substantially diminished."[212]

---

204. *Id.* at 647.

205. Musey, *supra* note 34, at 156 (quoting *Turner II*, 520 U.S. 180, 226 (1997) (Breyer, J., concurring) (quoting *Turner I*, 512 U.S. at 663)).

206. Musey, *supra* note 34, at 157 (footnote omitted).

207. *See Turner I*, 512 U.S. at 662 (discussing "the requirement of narrow tailoring").

208. *E.g.*, Musey, *supra* note 34, at 157–60 (questioning whether the *Turner II* decision would be followed if must-carry regulations are again challenged).

209. *Id.* at 158.

210. *Id.*

211. *Turner I*, 512 U.S. at 636.

212. Musey, *supra* note 34, at 158.

While must-carry privileges would be an optimal incentive for full relinquishment, allowing broadcasters to continue to reach a large majority of their audience without the need for spectrum, it may not survive a First Amendment attack.

Relocation Costs to Cover Leased-Access Fees (Cable). The Act provides significant protections for broadcasters choosing to relocate or share channels in the form of relocation cost payments.[213] Because Congress assumed broadcasters would not be "relocating," this protection is not extended to broadcasters choosing to fully relinquish.[214] Many stations that fully relinquish will successfully negotiate retransmission contracts with cable providers. However, some will not have the leverage to do so. As an option for those stations without the bargaining power, leased access to cable provides an appropriate alternative. As an added incentive to full relinquishment, the FCC could offer to pay leased-access rates for broadcasters choosing to fully relinquish as part of the broadcaster's "relocation costs."

Beginning in 1984, cable operators were required to allocate a significant percentage of their channels to commercial use by entities unaffiliated with the cable operator.[215] Cable operators with more than 100 activated channels are required to lease 15% of such channels.[216] The purpose of this requirement "is to promote competition in the delivery of diverse sources of video programming and to assure that the widest possible diversity of information sources are made available to the public from cable systems."[217] In 1992, Congress amended the requirements to set maximum prices for leased access and to regulate terms and conditions.[218] The history of mandatory leased-access price controls and regulations has been tumultuous, with programmers and cable operators disagreeing on a fair and consistent pricing scheme.[219] Currently, the maximum rate formula is "based on the 'average implicit fee' that non-leased access programmers are implicitly charged for carriage"—enough for the "cable operator to recover its costs and earn a profit."[220] The formula is complex and varies from

213. Middle Class Tax Relief and Job Creation Act of 2012, Pub. L. No. 112-96, § 6403(b)(4), 126 Stat. 156, 226–27 (to be codified at 47 U.S.C. § 1452).

214. *See supra* subpart III(C).

215. 47 U.S.C. § 532(b)(1) (2006).

216. *Id.* § 532(b)(1)(C).

217. *Id.* § 532(a).

218. Valuevision Int'l, Inc. v. FCC, 149 F.3d 1204, 1207 (D.C. Cir. 1998) (citing 47 U.S.C. § 532(c)(4)(A)).

219. *See id.* at 1206–08 (recounting FCC and statutory revisions to the mandatory leased-access pricing scheme and objections to those revisions by programmers and cable operators).

220. *Leased Access*, FCC ENCYCLOPEDIA (May 31, 2011), http://www.fcc.gov/encyclopedia/leased-access.

market to market, but it generally depends on subscriber revenue, programming costs, and the number of channels on an elected tier.[221]

In order to protect and incentivize broadcasters, the Act creates the TV Broadcaster Relocation Fund which will pay up to $1.75 billion for broadcaster relocation.[222] The FCC should interpret "relocation costs" to apply to broadcasters choosing to fully relinquish and use partial proceeds from the Relocation Fund to cover the price of leased access on cable channels. This increased incentive would accomplish multiple goals. First, it would be a strong incentive for many broadcasters to fully relinquish their 6 MHz allotment because they would continue to have the opportunity to program despite being off the air, while still receiving a significantly large reserve price in the auction. Second, it would reduce consumer harm because many of the local channels choosing to go off the air could still program on alternate media—which could be complemented with lifeline cable packages. Third, it would be consistent with leased access's stated purpose: to ensure diversity of information on local systems.

Finally, it may have an unforeseen positive externality. Cable companies are notorious for abusing their power to keep leased access expensive and difficult to obtain.[223] In 2008, the FCC released a report and order modifying leased-access rules, reducing the maximum permissible rate from an average implicit fee to the amount earned on the least profitable marginal channels.[224] The report was stayed by the Sixth Circuit.[225] Requiring the FCC to pay cable leased-access rates as part of their relocation-cost promise may incentivize the Commission to push for lower mandatory leased-access rates. This would increase access to other local programmers, many of whom never had licenses, to reach audiences via cable.

Payment of leased access to cable as an incentive for full relinquishment will benefit all constituencies as broadcasters remain on the air to serve their local communities, receive high auction proceeds, and free up spectrum for flexible licensing.

VI. Conclusion

The FCC and Congress have a unique opportunity to reallocate electromagnetic spectrum for flexible licensing, therefore remedying an

---

221. *Id.*

222. Middle Class Tax Relief and Job Creation Act of 2012, Pub. L. No. 112-96, § 6402, 126 Stat. 156, 224 (to be codified at 47 U.S.C. § 309(j)(8)(G)).

223. Bruce A. Olcott, Note, *Will They Take Away My Video-Phone if I Get Lousy Ratings?: A Proposal for a "Video Common Carrier" Statute in Post-Merger Telecommunications*, 94 COLUM. L. REV. 1558, 1577–78 (1994).

224. *In re* Leased Commercial Access, Report and Order and Further Notice of Proposed Rulemaking, 23 FCC Rcd. 2909, 2958–59 (Feb. 1, 2008).

225. *Leased Access*, *supra* note 220.

outdated and inefficient regulatory scheme. In order to do so, OTA TV broadcasters must agree to relinquish the airwaves on which they broadcast. While broadcasters may free up their spectrum by relocating to a shared channel or transferring to a new frequency, the full-relinquishment option of the Act's incentive auction plan is the optimal way for broadcasters to participate. The benefits of full relinquishment far outweigh the minimal harm to OTA TV viewers. Contrary to common perception, broadcasters who fully relinquish have ample opportunity to continue programming and thus serve local audiences. For smaller broadcasters without the ability to negotiate for retransmission or broadcasters reluctant to relinquish, providing incentives such as must-carry privileges or relocation compensation will ensure that broadcasters participate and audiences remain served. While the costs of these incentives are low, the benefits of clearing the airwaves are immeasurable.

—Michael Selkirk

* * *


TEXAS
TLR
LAW REVIEW