34:1 BERKELEY TECHNOLOGY LAW JOURNAL

2019

Pages

1

to

322

Berkeley Technology Law Journal

Volume 34, Number 1

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

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

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

BERKELEY TECHNOLOGY LAW JOURNAL

VOLUME 34 NUMBER 1 2019

TABLE OF CONTENTS

ARTICLES

SUBSCRIBER INFORMATION

The *Berkeley Technology Law Journal* (ISSN1086-3818), a continuation of the *High Technology Law Journal* effective Volume 11, is edited by the students of the University of California, Berkeley, School of Law and is published in print three times each year (March, September, December), with a fourth issue published online only (July), by the Regents of the University of California, Berkeley. Periodicals Postage Rate Paid at Berkeley, CA 94704-9998, and at additional mailing offices. POSTMASTER: Send address changes to Journal Publications, University of California, Berkeley Law—Library, LL123 Boalt Hall—South Addition, Berkeley, CA 94720-7210.

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

Subscriptions. Annual subscriptions are $65.00 for individuals and $85.00 for organizations. Single issues are $30.00. Please allow two months for receipt of the first issue. Payment may be made by check, international money order, or credit card (MasterCard/Visa). Domestic claims for non-receipt of issues should be made within 90 days of the month of publication; overseas claims should be made within 180 days. Thereafter, the regular back issue rate ($30.00) will be charged for replacement. Overseas delivery is not guaranteed.

Form. The text and citations in the *Journal* conform generally to the THE CHICAGO MANUAL OF STYLE (16th ed. 2010) and to THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (Columbia Law Review Ass'n et al. eds., 20th ed. 2015). Please cite this issue of the *Berkeley Technology Law Journal* as 33 BERKELEY TECH. L.J. ___ (2019).

BTLJ ONLINE

The full text and abstracts of many previously published *Berkeley Technology Law Journal* articles can be found at http://www.btlj.org. Our site also contains a cumulative index; general information about the *Journal*; the BTLJ Blog, a collection of short comments and updates about new developments in law and technology written by BTLJ members; and *BTLJ Commentaries*, an exclusively online publication for pieces that are especially time-sensitive and shorter than typical law review articles.

INFORMATION FOR AUTHORS

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

Format. Submissions are accepted in electronic format through Scholastica online submission system. Authors should include a curriculum vitae and resume when submitting articles, including his or her full name, credentials, degrees earned, academic or professional affiliations, and citations to all previously published legal articles. The Scholastica submission website can be found at https://scholasticahq.com/law-reviews.

Citations. All citations should conform to THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (Columbia Law Review Ass'n et al. eds., 20th ed. 2015).

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

Donors

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

Partners

Cooley LLP

Fenwick & West LLP

Hogan Lovells

Orrick, Herrington & Sutcliffe LLP

White & Case LLP

Benefactors

Baker Botts LLP

Covington & Burling LLP

Fish & Richardson P.C.

Jones Day

Kirkland & Ellis LLP

Latham & Watkins LLP

McDermott Will & Emery

Morrison & Foerster LLP

Polsinelli LLP

Sidley Austin LLP

Weil, Gotshal & Manges LLP

Wilmer Cutler Pickering Hale and Dorr LLP

Wilson Sonsini Goodrich & Rosati

Winston & Strawn LLP

Corporate Benefactors

Atlassian

Computer & Communications Industry Association

Cornerstone Research

Future of Privacy Forum

Google, Inc.

Hewlett Foundation, through the Center for Long-term Cybersecurity

Intel

InventionShare

LitiNomics

Microsoft Corporation

Mozilla

NERA Economic Consulting

Nokia

Palantir

RLM Trialgraphix

The Walt Disney Company

Members

Baker & McKenzie LLP

Crowell & Moring

Desmarais LLP

Durie Tangri LLP

Finnegan, Henderson, Farabow, Garrett & Dunner, LLP

GTC Law Group LLP & Affiliates

Haynes and Boone, LLP

Hickman Palermo Becker Bingham

Irell & Manella LLP

Keker Van Nest & Peters LLP

Kilburn & Strode

Kilpatrick Townsend & Stockton LLP

Knobbe Martens LLP

Kwan & Olynick LLP

Morgan, Lewis & Bockius LLP

Paul Hastings LLP

Robins Kaplan LLP

Ropes & Gray LLP

Simpson Thacher & Bartlett LLP

Troutman Sanders LLP

Van Pelt, Yi & James LLP

Weaver Austin Villeneuve & Sampson LLP

Board of Editors 2018–2019

Membership

Vol. 34 No. 1

Associate Editors

Members

Trenton Davis
Liz Douglass
Ida Ebeid
Rucha Ekbote
Katelyn Feliciano
Olgamaris Fernandez
Moritz Flechsenhar
Yesenia Flores
Jason Francis
Logan Freeborn
Ravin Galgotia
Maribel Garcia
Garima Garg
Kerensa Gimre
Michael Kostukovsky
Kristina Krasnikova
Soumya Jogaiah Krishnaraju
Emma Lee
Jian Lee
Killian Lefevre
Elly Leggatt
Xiaocao Li
Ashleigh Lussenden
Aartika Maniktala
Marissa Medansky
Alex Milne
Alexandre Mochon
Schuyler Standley
Lyric Stephenson
Daniel Todd
Viviane Trojan
Daniel Twomey
Edgar Vega
Yuhan Wang
Grace Winschel
Jolene Xie
Kevin Yang
Alison Yardley
Clark Zhang
Jieyu Zhang
Pengpeng Zhang
Evan Zimmerman

BTLJ ADVISORY BOARD

BERKELEY CENTER FOR LAW & TECHNOLOGY 2018–2019

Executive Director

JIM DEMPSEY

Faculty Directors

KENNETH A. BAMBERGER
CATHERINE CRUMP
CATHERINE FISK
CHRIS HOOFNAGLE
SONIA KATYAL

PETER S. MENELL
ROBERT P. MERGES
DEIRDRE K. MULLIGAN
TEJAS NARECHANIA
ANDREA ROTH

PAMELA SAMUELSON
PAUL SCHWARTZ
ERIK STALLMAN
JENNIFER M. URBAN
MOLLY S. VAN HOUWELING

Fellow

KATHRYN HASHIMOTO

Staff Directors

JANN DUDLEY
RICHARD FISK

IRYS SCHENKER
MATTHEW RAY

GRANTS

W. Nicholson Price II[†]

ABSTRACT

Innovation is a primary source of economic growth and is accordingly the target of substantial academic and government attention. Grants are a key tool in the government's arsenal to promote innovation, but legal academic studies of that arsenal have given them short shrift. Although patents, prizes, and regulator-enforced exclusivity are each the subject of substantial literature, grants are typically addressed briefly, if at all. According to the conventional story, grants may be the only feasible tool to drive basic research, as opposed to applied research, but they are a blunt tool for that task.

Three critiques of grants underlie this narrative: grants are allocated by government bureaucrats who lack much of the relevant information for optimal decision-making; grants are purely ex ante funding mechanisms and therefore lack accountability; and grants misallocate risk by saddling the government all the downside risk and giving the innovator all the upside. These critiques are largely wrong. Focusing on grants awarded by the National Institutes of Health (NIH), the largest public funder of biomedical research, this Article delves deeply into how grants actually work. It shows that—at least at the NIH—grants are awarded not by uninformed bureaucrats, but by panels of knowledgeable peer scientists with the benefit of extensive disclosures from applicants. It finds that grants provide accountability through repeated interactions over time. And it argues that the upside of grant-investments to the government is much greater than the lack of direct profits would suggest.

Grants also have two marked comparative strengths as innovation levers: they can support innovation where social value exceeds appropriable market value, and they can directly support innovation enablers—the people, institutions, processes, and infrastructure that shape and generate innovation. Where markets undervalue some socially important innovations, like

DOI: https://doi.org/10.15779/Z38VQ2SB0K

[†] Assistant Professor of Law, University of Michigan Law School. JD, Columbia Law School, 2011. PhD (Biological Sciences), Columbia University, 2010. For helpful comments and conversations, I thank Sam Bagentsos, Nick Bagley, Ana Bracic, Sarah Burstein, Jorge Contreras, Rebecca Eisenberg, Paul Gugliuzza, Daniel Hemel, Jim Hines, Camilla Hrdy, Tabrez Ibrahim, Peter Karol, Dmitry Karshtedt, Jessica Litman, Kyle Logue, Lisa Larrimore Ouellette, Laura Pedraza-Fariña, Gabriel Rauterberg, Alex Roberts, Ana Santos Rutschman, Rachel Sachs, Margo Schlanger, Jake Sherkow, Ofer Tur-Sinai, and Patti Zettler. Erin Edgerton, Jake Flood, Ana Kolosova, and Jonathan Tietz provided excellent research assistance. This work benefited from feedback at the RIBS Junior Scholars Workshop at Michigan Law School, the Health Law Professors' Conference at Georgia State University, the Wiet Life Sciences Law Conference at Loyola Law School Chicago, the University of New Hampshire's IP Scholars Roundtable, the Junior IP Scholars Association's Gnocchi Workshop at the University of Oklahoma, and the Works in Progress in Intellectual Property conference at Case Western Reserve University. All errors are my own.

cures for diseases of the poor, grants can help. Grants can also enable innovation by supporting its inputs: young or exceptional scientists, new institutions, research networks, and large datasets. Taken as a whole, grants do not form a monolithic, blunt innovation lever; instead, they provide a varied and nuanced set of policy options. Innovation scholars and policymakers should recognize and develop the usefulness of grants in promoting major social goals.

TABLE OF CONTENTS

I. INTRODUCTION

Grants play a key role in innovation policy. The federal government spent over $64 billion in 2016 in grants to support scientific research.[1] That sum is vastly more than the government spends on prizes (under $0.1 billion), nearly

[1] *Historical Trends in Federal R&D*, AM. ASS'N FOR ADVANCEMENT SCI. https://www.aaas.org/programs/r-d-budget-and-policy/historical-trends-federal-rd [https://perma.cc/3P22-A45C] (last visited Mar. 9, 2019).

an order of magnitude greater than what it spends on research and development tax credits (about $10 billion), and comparable to what it spends on patents through a shadow tax on consumers (between $30 and $700 billion, though difficult to estimate).[2] Grants are especially prominent in the life sciences. The National Institutes of Health (NIH) is the world's largest public funder of biomedical research.[3] Every year, it administers over $29 billion in grant funding to over 300,000 researchers in over 2,500 institutions.[4] Through their scale and ubiquity, grants significantly shape the progress of science and innovation. Grants help determine which areas of science are studied and how, make or break the careers of academic and non-academic scientists alike, and guide the creation of new institutes and discipline-spanning resources.

So how should the grant system operate? When should we deploy grants instead of patents or prizes to drive innovation? Whom should we fund and what policies should govern that funding? These questions are not rhetorical: 2017 saw a high-profile fight between the Trump Administration and Congress about science funding levels[5] and an intense discussion in the scientific community about new NIH grant-funding policies.[6]

If these questions addressed changes to patent law, we could draw on an extensive literature about how patents shape innovation, what changes would have what impacts, and what we should think about when proposing new

[2] Daniel J. Hemel & Lisa Larrimore Ouellette, *Beyond the Patents–Prizes Debate*, 92 TEX. L. REV. 303, 361, 371 (2013) (defining grants as including both funds directed to external researchers and funds spent on direct government research and basing patent expenditures on the patent-enabled supra-competitive pricing that constitutes a "shadow tax" on consumers of the patented good).

[3] *See Grants & Funding*, NIH, https://www.nih.gov/grants-funding [https://perma.cc/8BJY-AZ4D] (last visited Mar. 9, 2019).

[4] *See Budget*, NIH, https://www.nih.gov/about-nih/what-we-do/budget [https://perma.cc/PKP5-4WVZ] (last visited Mar. 9, 2019) [hereinafter NIH, *Budget*].

[5] *See, e.g.*, Joel Achenbach & Lena H. Sun, *Trump Budget Seeks Huge Cuts to Science and Medical Research, Disease Prevention*, WASH. POST (May 23, 2017), https://www.washingtonpost.com/news/to-your-health/wp/2017/05/22/trump-budget-seeks-huge-cuts-to-disease-prevention-and-medical-research-departments/ [https://perma.cc/UAY9-28Y5] (noting the early unfavorable reactions to Trump's proposed budget); Robert Pear, *Congress Rejects Trump Proposals to Cut Health Research Funds*, N.Y. TIMES (Sept. 11, 2017), https://www.nytimes.com/2017/09/11/us/politics/national-institutes-of-health-budget-trump.html [https://perma.cc/KA32-BCY3] (noting that Congress rejected Trump's proposed budget and introduced a bipartisan bill to increase spending).

[6] *See, e.g.*, *Develop Your Budget*, NIH, https://grants.nih.gov/grants/how-to-apply-application-guide/format-and-write/develop-your-budget.htm [https://perma.cc/6B4P-EFPK] (last visited Oct. 31, 2018) (providing instructions to create a budget and noting that there are "spending caps on certain expenses" in addition to salary caps); Sara Reardon, *NIH Announces Grant Limits*, 545 NATURE 142 (2017) (discussing the concerns of the scientific community in response to the NIH's new budget policy).

policies.[7] If these questions considered the structure or funding of prizes for achieving innovation goals, we could reach for another extensive literature tackling similar issues.[8] And if we wished to debate the relative merits of patents, prizes, pure market allocation, government procurement, tax subsidies for research-and-development, and grants, a substantial volume of scholarship addresses such comparative issues.[9] But the grant system itself? That occupies a much emptier shelf in the library of innovation law.[10]

In the uncommon instances where grants appear in this literature, they appear in comparative work evaluating the advantages and disadvantages of different policy mechanisms for promoting innovation. In this context, a consistent argument holds that grants suffer from an information disadvantage relative to patents, and, to a lesser extent, prizes and tax incentives, because they do not effectively aggregate private information.[11] A closely related point is that grants are particularly useful at funding basic research—that is, early-

[7] *See, e.g.*, Robert P. Merges & Richard R. Nelson, *On the Complex Economics of Patent Scope*, 90 COLUM. L. REV. 839 (1990); Dan L. Burk & Mark A. Lemley, *Policy Levers in Patent Law*, 89 VA. L. REV. 1575 (2003); Edmund W. Kitch, *The Nature and Function of the Patent System*, 20 J.L. & ECON. 265 (1977); Ian Ayres & Paul Klemperer, *Limiting Patentees' Market Power Without Reducing Innovation Incentives: The Perverse Benefits of Uncertainty and Non-Injunctive Remedies*, 97 MICH. L. REV. 985 (1999); Stuart J.H. Graham et al., *High Technology Entrepreneurs and the Patent System: Results of the 2008 Berkeley Patent Survey*, 24 BERKELEY TECH. L.J. 1255 (2009); John M. Golden, *Principles for Patent Remedies*, 88 TEX. L. REV. 505 (2010); Craig A. Nard & John F. Duffy, *Rethinking Patent Law's Uniformity Principle*, 101 NW. U. L. REV. 1619 (2007); Colleen V. Chien, *From Arms Race to Marketplace: The Complex Patent Ecosystem and Its Implications for the Patent System*, 62 HASTINGS L.J. 297 (2010); Benjamin N. Roin, *The Case for Tailoring Patent Awards Based on Time-to-Market*, 61 UCLA L. REV. 672 (2014).

[8] *See* Benjamin N. Roin, *Intellectual Property Versus Prizes: Reframing the Debate*, 81 U. CHI. L. REV. 999, 1003–05 (2014) (noting that "the past two decades have seen a virtual explosion of scholarship on prize systems, particularly within the economic and legal literatures on intellectual property, but also in political philosophy and public health" and providing extensive citations).

[9] *See* Hemel & Ouellette, *supra* note 2, at 305 ("In recent years, articles comparing the relative merits of patents, prizes, and grants have consumed thousands of pages in law reviews and economics journals.") (citing Peter S. Menell & Suzanne Scotchmer, *Intellectual Property Law*, *in* 2 HANDBOOK OF LAW AND ECONOMICS 1473, 1530–34 (A. Mitchell Polinsky & Steven Shavell eds., 2007) (reviewing recent literature)).

[10] *See, e.g.*, Laura Pedraza-Fariña, *The Social Origins of Innovation Failures*, 70 SMU L. REV. 377, 443 (2017) ("Legal scholarship on intellectual property and innovation law more broadly has paid comparatively little attention to how to design grants and prizes to foster innovation, and how grant-making interacts with other innovation policies—and patents in particular.").

[11] *See* Harold Demsetz, *Information and Efficiency: Another Viewpoint*, 12 J.L. & ECON. 1, 11–14 (1969). This argument applies with equal force to other exclusivity-based incentive mechanisms, such as trade secrecy or regulatory exclusivity, since all exclusivity mechanisms rely on allowing the innovator to charge supra-competitive prices to capture a greater portion of the social welfare benefits of an innovation.

stage research without immediate commercial applications—because firms tend to undervalue basic research, which has substantial positive knowledge externalities.[12]

Within the innovation law literature's relatively sparse descriptions of grants, three critiques recur—sometimes as explicit critiques, sometimes as assumptions, sometimes as characterizations—about flaws in the grant system. To be clear, not all scholars writing about grants raise all these critiques, or make them uncritically. In this literature, grants are undertheorized, which is both the point and the challenge. I reviewed closely the existing, brief discussions of grants in the law-and-innovation literature, and common threads emerged.

Part II describes these critiques. First, grants are allocated by government bureaucrats who lack the market-value knowledge possessed by private firms and therefore make suboptimal decisions about allocating funding to projects.[13] Second, because grants provide non-contingent ex ante funding, they lack accountability and thus cannot ensure efficient and hard work by innovators.[14] And third, grants allocate risk suboptimally: the grantor takes essentially all of the downside risk of the project (if the innovation fails, the government is still out the money with nothing to show for it) and receives little of the upside benefit (if the project succeeds, the innovator licenses or commercializes the innovation, while the government misses out on the profits and may even end up paying high prices for the innovation).[15] Taken together, these critiques lead to the conclusion that while grants may be an adequate, if rather blunt, tool to drive basic research for which other innovation levers are unhelpful, those other levers are often preferable when available. Jonathan Adler, for instance, actively critiques the grant system on these grounds, concluding that "the federal government should shift a substantial portion of climate-related research and development funding from grants to prizes."[16] I suspect that these critiques are also responsible for the relative dearth of scholarship examining grants in depth. If grants are generally viewed as good for basic research but flawed relative to other incentive levers, why spend much time thinking about them?[17]

[12] *See infra* Part II.

[13] *See infra* Section II.A.

[14] *See infra* Section II.B.

[15] *See infra* Section II.C.

[16] Jonathan H. Adler, *Eyes on a Climate Prize: Rewarding Energy Innovation to Achieve Climate Stabilization*, 35 HARV. ENVTL. L. REV. 1, 4 (2011); *see infra* Part II.

[17] There are other potential explanations. Laura Pedraza-Fariña and Stephanie Bair, for example, argue that legal scholars of innovation have focused on solving the free-rider problem to the exclusion of other innovation challenges. *See* Stephanie Bair & Laura Pedraza-

The reality of the current grant system belies these three critiques. Part III describes the grant system as it functions today, with substantial emphasis on grants awarded by the NIH—perhaps the world's most prominent grant funder—to researchers at other institutions, and rebuts each critique.[18] First, the mechanics of grant application, review, and funding refute the narrative that grants are allocated by information-poor bureaucrats. The grant system uses a rigorous process of peer review to determine which proposals will be funded. Part of this process involves detailed applications, which requires potential grant recipients to share their own private information about the likely costs and potential value of the proposed research. The evaluation itself leverages the expertise of scientists with relevant experience and knowledge. And the entire process is coordinated by agency representatives who combine their own scientific background with knowledge about the innovation priorities of the NIH and the government more generally.

Second, grantees are in fact accountable for grant-funded research. Each grant operates within a context of ongoing funding streams, reporting obligations, and repeat players. Even though any individual grant may lack its own strong accountability mechanisms, the practical need to get the next grant creates accountability for grant recipients.[19]

Third, the government gets more out of grants than the risk-allocation critique implies. It's true that the government does not usually profit directly from grant-funded innovations, whether they succeed or fail. But the government realizes a wide range of social benefits from innovation efforts, including the creation of negative knowledge, the generation of innovation structures, and the development of human capital.

Mistaken assumptions or inaccurate critiques change the relative desirability of grants as a substitute for other innovation levers when those levers fail. Consider patentable subject matter. Between 2012 and 2014, the Supreme Court held unpatentable a broad swath of inventions that could be

Fariña, *Anti-Innovation Norms*, 112 NW. L. REV. 1069, 1076–78 (2018); *see also* Joshua D. Sarnoff, *Government Choices in Innovation Funding (with Reference to Climate Change)*, 62 EMORY L.J. 1087, 1100 (2013) (similarly lamenting the narrow focus of legal-academic literature). Because grants do not address free-rider critiques directly, they may be of less interest to legal scholars with that focus.

[18] I argue that basic lessons from the NIH are generalizable, *see infra* note 203 and accompanying text, but even to the extent they are not, understanding the workings of the world's largest public funder of biomedical research provides useful insight. *See* NIH, *supra* note 3.

[19] The ongoing grant cycle has other benefits. For instance, the ongoing need to seek future grants impels grant recipients to generate publications that disclose results of funded work.

characterized as "laws of nature, natural phenomena, [or] abstract ideas."[20] These decisions prompted scholarly outcry: among other issues, what incentives would remain for inventions that subject to this characterization, like medical diagnostic methods or human genetic tests?[21] In fact, the Court raised exactly this question at oral argument.[22] As it turns out, many medical diagnostics and human genetic tests have been developed in large part by grant-funded researchers. Rather than worrying about decreased patent incentives, perhaps Congress should increase grant funding for these inventions instead.[23] If we think grants are fundamentally flawed innovation levers, they are less likely to seem like good substitutes when other levers fail. If, on the contrary, we are to use grants appropriately as a part of the innovation toolbox, we should know how they really work: when they are preferable substitutes, when they work poorly, and when they work best in concert with other innovation incentives.[24]

Part IV describes the rich tools the grant system supplies to policymakers, focusing on grants' two key comparative advantages. First, grants can support innovations whose social value exceeds their appropriable market value. This describes basic research; because later applications of basic research are variable and unpredictable, it has substantial spillovers (positive knowledge externalities), and is undersupplied by private firms relative to its social benefit.[25] Private firms also generate inadequate information about which basic research is worth funding. But a panel of experienced peer reviewers,

[20] Mayo Collaborative Servs. v. Prometheus Labs., Inc., 566 U.S. 66, 66 (2012) (citing Diamond v. Diehr, 450 U.S. 175, 185 (1981)); *see also* Ass'n for Molecular Pathology v. Myriad Genetics, Inc., 569 U.S. 576, 589 (2013); Alice Corp. Pty. Ltd. v. CLS Bank Int'l, 573 U.S. 208, 134 S. Ct. 2347, 2354 (2014).

[21] *See* Rachel E. Sachs, *Innovation Law and Policy: Preserving the Future of Personalized Medicine*, 49 U.C. DAVIS L. REV. 1881, 1907–13 (2016) (discussing the difficulties of obtaining patents in diagnostic methods); Rebecca S. Eisenberg, *Diagnostics Need Not Apply*, 21 B.U. J. SCI. & TECH. L. 256, 264–78 (2015) (discussing how diagnostic methods have been categorized as "natural laws" rather than "applications").

[22] *See* Lisa Larrimore Ouellette, *Patentable Subject Matter and Nonpatent Innovation Incentives*, 5 U.C. IRVINE L. REV. 1115, 1116–17 (2015) (citing Justice Sotomayor's questions during oral argument of *Myriad* and *Mayo*).

[23] *See id.* at 1137–41 (discussing other incentives that can take the place of absent patent incentives).

[24] No innovation lever stands on its own; an innovation may be grant-funded in early phases, patented shortly thereafter, developed using secret processes and relying on tax-incentives, and even win a prize at the end. *See, e.g.*, Pierre Azoulay et al., *Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules*, 86 REV. ECON. STUD. 117, 140 (2019) (finding that a $10 million boost in NIH funding leads to around 2.5 additional patents).

[25] *See generally* Richard R. Nelson, *The Simple Economics of Basic Scientific Research*, 67 J. POL. ECON. 297, 302–04 (1959).

combined with disclosures from grant-seeking researchers, may be able to make precisely that determination. At a broader level, the market systematically undervalues some forms of innovation because market demand does not reflect social welfare value. A powerful example is innovation targeting diseases of the poor; because the poor often cannot pay for drugs, market signals do not reflect the social welfare benefits of developing those drugs. The grant system's reliance on non-price signals brings risks of inefficiency or cronyism, but its incorporation of non-market information also allows different, useful allocation of funds beyond what markets would pick.

Second, grants can directly support innovation enablers—the people, institutions, processes, and infrastructural resources involved in innovation—in a way largely unavailable to other forms of directed innovation incentives, especially patents and prizes. Basic research serves this role when it provides the grounding for later research, but it is only one example. Grants can develop human capital by providing training or otherwise enabling the research of young scientists who will have longer careers ahead of them. Grants can also target the processes or institutions of innovation by providing resources specifically for interdisciplinary research (to build collaborations and boundary-crossing networks) or for institutions (to provide physical or other resources for collections of individuals). Finally, they can support infrastructure, including datasets that enable future innovation, such as the Precision Medicine Initiative's All of Us dataset or the Human Genome Project (both NIH-funded).

When policymakers can leverage the grant system's strengths, grants can be an effective innovation lever. But the inverse is also true. In situations where private, market-based information accurately reflects the social value of an innovation, grants are probably not the best lever to drive that innovation because that private information can lead to an efficient allocation of innovative activity among firms and innovation targets.

This Article argues that the dominant picture of scientific grants in the innovation literature—the picture of a relatively straightforward and flawed tool mostly good for basic research—is far too simple. Grants form their own complex, massive set of innovation tools, with their own comparative strengths, and are a far larger, better, and more varied part of the innovation system than the innovation law literature has recognized.

II. GRANTS IN THE INNOVATION LAW LITERATURE

Grants are undertheorized in the legal innovation literature. Where they appear, it is principally as part of a comparison with other sorts of innovation incentives, though even those comparisons tend to focus on patents and prizes, rather than grants.[26] Daniel Hemel and Lisa Ouellette, for instance, compare the innovation incentives of patents, prizes, grants, and tax R&D incentives.[27] They group incentives along three axes—who decides what innovation will be funded, who pays for the innovation, and when is the innovation funded—and conclude that each incentive is useful at different times.[28] Grants, they suggest, are most effective when the government is especially good at identifying costs and benefits and when social benefits exceed market signals of value—one of the two key strengths I describe here.[29] They also note an important timing feature of grants: ex ante funding can enable otherwise capital-constrained entities to innovate.[30] Joshua Sarnoff, Brett Frischmann, and Jonathan Adler have also considered grants in comparisons of innovation levers.[31] Characterizations of grants as an innovation incentive, whether comparative or otherwise, have tended to emphasize the information disadvantage faced by the grant system, but also the positive role of grants in funding basic research.

The basic information-asymmetry story proceeds as follows. Innovators determine whether to invest in a particular innovation based on their private

[26] *Compare* Roin, *supra* note 8, at 1001–06 (providing approximately 4 pages worth of citations on prizes versus the patent system), *with* Hemel & Ouellette, *supra* note 2, at 320–21 (citing, in a prominent and thorough taxonomy of innovation incentives, only one unpublished manuscript and one law review article partially focused on grants). Camilla Hrdy has briefly addressed grants in the context of analyzing federal versus state and local incentives for innovation. *See* Camilla Hrdy, *Patent Nationally, Innovate Locally*, 31 BERKELEY TECH. L.J. 1301, 1357–63 (2016) [hereinafter Hrdy, *Patent Nationally*] (discussing federal financing for innovation, including grants, and arguing that such funding is limited to research with national benefits); *see also* Camilla Hrdy, *Commercialization Awards*, 2015 WIS. L. REV. 13, 52–53 (2015) [hereinafter Hrdy, *Commercialization Awards*] (discussing Small Business Innovation Research (SBIR) awards granted by federal research agencies like NIH).

[27] Hemel & Ouellette, *supra* note 2, at 310–15.

[28] *See id.* at 326–52.

[29] *See id.* at 375–76. These two features are both involved in the grant system's ability to use different information than markets, described in Part IV.

[30] *Id.* at 334–38.

[31] *See* Sarnoff, *supra* note 17, at 1089–90 (considering a broad range of potential incentives in the climate-change context and noting the lack of empirical information on grant functioning); Joshua D. Sarnoff, *The Likely Mismatch Between Federal Research & Development Funding and Desired Innovation*, 18 VAND. J. ENT. & TECH. L. 363, 372 (2016) (lamenting the focus of innovation law scholarship on intellectual property and market solutions to innovation) [hereinafter Sarnoff, *Likely Mismatch*]; Brett Frischmann, *Innovation and Institutions: Rethinking the Economics of U.S. Science and Technology Policy*, 24 VT. L. REV. 347, 352–53, 356, 389–90 (1999) (noting that grants are useful for the production of public goods but tax incentives are preferable in other situations); Adler, *supra* note 16, at 3–4 (comparing grants and prizes in the context of climate change technology and concluding that prizes are generally superior).

estimations of the cost of innovating and the innovation's market value.[32] Non-grant mechanisms alter this private-information-based calculus: patents allow firms to capture a larger fraction of the expected value of the innovation,[33] prizes typically set a known reward against the privately-estimated cost of innovating,[34] and tax incentives directly defray the cost of innovating.[35] Grants, on the other hand, provide ex ante funds to pay innovation costs directly and do not leverage private estimations of market value.

In 1983, Brian Wright showed formally that for patents to be superior to other innovation incentives, private firms must have more information than government funders.[36] Scholars tend to agree that private firms have such an information advantage.[37] Suzanne Scotchmer and Nancy Gallini, for instance, built on Wright's analysis and noted that grants are poor aggregators of private information.[38]

However, scholars have also long agreed that grants are important for funding basic research, though this agreement is grounded in the economics literature rather than the legal literature.[39] Basic research is aimed at increasing our scientific understanding of the world rather than focusing on useful products. In 1959, Richard Nelson noted that basic research has potential innovation benefits across a wide range of outputs and is often highly risky.[40] As a result, private industry tends to invest in basic research at socially suboptimal levels.[41] Kenneth Arrow reiterated this argument in 1962 and suggested that government funding of innovation helps resolve the problem, though such funding raises questions of how much to spend and how to

[32] Hemel & Ouellette, *supra* note 2, at 326–27.

[33] *See id.* at 327–28.

[34] *See id.* at 327.

[35] *See id.* at 328.

[36] *See* Brian D. Wright, *The Economics of Invention Incentives: Patents, Prizes, and Research Contracts*, 73 AM. ECON. REV. 691, 691 (1983). Among other things, Wright deliberately omits the possibility that the modeled innovation would provide information useful for future innovations and therefore of independent social value. *See id.* at 692 n.1.

[37] *See, e.g.*, Hemel & Ouellette, *supra* note 9, at 327 ("Patents' ability to take advantage of private information is well recognized in the innovation-policy literature.").

[38] Nancy Gallini & Suzanne Scotchmer, *Intellectual Property: When Is It the Best Innovation System?*, 2 INNOV. POL'Y & ECON. 51, 54, 55–57 (2002).

[39] *See, e.g.*, Joseph E. Stiglitz, *Economic Foundations of Intellectual Property Rights*, 57 DUKE L.J. 1693, 1721 (2008) (noting that grants "are probably the most important component of the innovation system, in supporting basic research"); *id.* at 1724 (claiming general agreement that grants are the right incentive for basic research, and that the only debate is about applied research).

[40] *See* Nelson, *supra* note 25, at 304.

[41] *See id.*

allocate it.[42] Despite these questions of allocation raised by Arrow, other incentive mechanisms, including patents and trade secrecy, are poor drivers of the production of basic knowledge, giving grants the comparative advantage.[43]

These assessments of grants, especially in comparison with other innovation policy levers, frequently incorporate three substantive critiques about how grants work. First is reliance on decision-making by government bureaucrats who often lack market actors' superior knowledge; second is the loss of accountability and incentives because grants rely on purely ex ante funding; and third is problematic risk-allocation because the funder bears the entire downside risk of the project and captures little of the upside benefit. Some find these critiques essentially dispositive; Jonathan Adler concludes that while "[f]ederal funding of science is worthwhile, particularly for basic scientific research[,] federal R&D money rarely produces commercially viable technologies or dramatic technological innovation."[44] The following Sections detail each critique.

A. Bureaucratic Decision-Making

Some criticize the grant system because it puts funding decisions in the hands of relatively uninformed government bureaucrats. As Adler puts it, "With government research grants . . . a federal agency typically determines the goal to be achieved, the means to achieve that goal, and who will receive funding to pursue it."[45] Frischmann agrees: "[T]he selection process for grants relies on the government's ability to assess the desirability of a project when compared with an array of others . . . ."[46] Lobbying groups have sometimes seized on this complaint; the director of the Traditional Values Coalition described NIH funding as "nameless, faceless bureaucrats doling out money

[42] Kenneth J. Arrow, *Economic Welfare and the Allocation of Resources for Invention*, *in* THE RATE AND DIRECTION OF INVENTIVE ACTIVITY: ECONOMIC AND SOCIAL FACTORS 609, 619, 623 (1962).

[43] *See, e.g.*, Amy Kapczynski & Talha Syed, *The Continuum of Excludability and the Limits of Patents*, 122 YALE L.J. 1900, 1905–06 (2013) (noting the challenge of appropriating the benefits of basic knowledge); Eisenberg, *supra* note 21, at 256 (noting the inability of patents to claim basic biomedical knowledge used in diagnostics under current law); Peter Lee, *Social Innovation*, 92 WASH. U. L. REV. 1, 24–42 (2014) (describing limitations of patents in creating incentives for social innovation); *id.* at 47–52 (describing how government grants might help create such incentives).

[44] Adler, *supra* note 16, at 30.

[45] *Id.* at 14.

[46] Frischmann, *supra* note 31, at 353; *see* Hemel & Ouellette, *supra* note 2, at 307 ("For grants . . . the government tailors the reward on a project-by-project or discovery-by-discovery basis.").

like a federal ATM . . . ."[47]

This critique can involve concerns of either inadequate information or cronyism. First, the government lacks information, at least relative to firms. Firms may have private knowledge about both the general costs and benefits of a potential innovation (relevant to the choice of which innovation to fund) and about their own costs in pursuing that innovation (relevant to the choice of which firm should pursue the innovation); the patent system especially leverages private knowledge by letting firms decide which innovation to pursue.[48] Conversely, the government's lack of this private information likely leads to suboptimal choices about what innovation to fund and who should undertake it.[49] Michael Abramowicz argues specifically in the context of orphan drugs that government officials are ill-equipped to distinguish efficient from inefficient innovation subsidies.[50] Zachary Liscow and Quentin Karpilow capture this general concern about information asymmetries when they note IP scholars' deep "skepticism toward the government 'picking winners' to encourage innovation in some technologies over others."[51]

Second, leaving funding decisions in the hands of bureaucrats may result in cronyism, favoritism, and political pressure shaping the process of grant-funding and scientific progress. Adler argues that historically, patrons of

[47] Rick Weiss, *NIH Faces Criticism on Grants*, WASH. POST (Oct. 30, 2003), https://www.washingtonpost.com/archive/politics/2003/10/30/nih-faces-criticism-on-grants/504677ed-4c30-498e-b458-c992ecf6c6f4/?utm_term=.df4269595c8d [https://perma.cc/65YK-JZRT]. The Coalition's concerns eventually led to Senate hearings. Rick Weiss, *Critics of NIH Studies Prompt Senate Hearings*, WASH. POST (Jan. 19, 2004), https://www.washingtonpost.com/archive/politics/2004/01/19/critics-of-nih-studies-prompt-senate-hearing/fa9de180-39ab-4dca-8757-34e7dfb80b4e/?utm_term=.4a2e01a478c7 [https://perma.cc/2UHY-J9KX].

[48] *See* Gallini & Scotchmer, *supra* note 38, at 54–55 (explaining that IP has substantial benefits if firms have superior knowledge); *see also* Wright, *supra* note 36, at 703 (noting that patents benefit from "ex ante researcher information relating to the value of the invention").

[49] *See* Adler, *supra* note 16, at 29 ("Allocating grant money effectively requires the grant-making entity to pick 'winners' and 'losers,' something the government has rarely done well."). Frischmann notes:

> [T]he selection process for grants relies on the government's ability to assess the desirability of a project when compared with an array of others . . . . If the research is expected to further a commercial end then tax incentives may be more effective than grants because final project selection is left to the best informed investor, the firm.

Frischmann, *supra* note 31, at 353.

[50] *See* Michael Abramowicz, *Orphan Business Models: Toward a New Form of Intellectual Property*, 124 HARV. L. REV. 1362, 1366–67 (2011).

[51] Zachary D. Liscow & Quentin Karpilow, *Innovation Snowballing and Climate Law*, 95 WASH. U. L. REV. 387, 390 n.9 (2017); *see also* Lee, *supra* note 43, at 52 ("[G]overnments are notoriously poor at 'picking winners.' ").

science preferred grants to prizes because grants entailed greater discretion, so that patrons could "reward their friends and allies and ensure that only those with the right ideas received funding."[52] Cronyism and corruption lead to many ills, including inefficiency, decreased trust in government, and lower innovation, as only ideas that match the idiosyncratic preferences of the funder receive funding.

B. Unaccountable Ex Ante Incentives

A second major critique relates to the ex ante nature of grant funding and its consequent lack of accountability. Grants provide funds ex ante to researchers without conditioning the funds on success.[53] Thus, the argument goes, grants provide less accountability and lower incentives to researchers to work hard and to use resources efficiently.[54] Sarnoff laments that "direct subsidies may be provided to university professors who fail to produce quality research" and thus "over-reward innovation efforts."[55] As Gallini and Scotchmer memorably describe it, in one-off grant contexts, "researchers might be inclined to 'take the money and run.' "[56] Hemel and Ouellette add that this unconditionality may cause problems earlier in the process, leaving grant-seeking researchers with lower incentives when choosing projects.[57]

The researcher, in this critique, has little skin in the game, in striking contrast to patents, prizes, or even R&D tax incentives. Under those regimes, the researcher must spend her own money to conduct the research or acquire funding from private sources with, presumably, strings attached.[58] And if she

[52] Adler, *supra* note 16, at 23 (citing Robin Hanson, Patterns of Patronage: Why Grants Won over Prizes in Science 17 (July 28, 1998) (unpublished manuscript) (on file with Harvard Law School Library)); *see also id.* at 29 ("[T]raditional grant funding is more subject to political pressure[.]").

[53] Indeed, if grants were conditioned on success, they would merely be prizes with precedent loans. Grants may condition continued funding on other requirements, such as continued reporting, documented expenditures, or something else; these complications will be described below. *See infra* Section III.B.2.a.

[54] *See, e.g.*, Arrow, *supra* note 42, at 624 (noting this problem and describing potential mitigating factors); *see also* Sarnoff, *supra* note 17, at 1125.

[55] Sarnoff, *supra* note 17, at 1125.

[56] Gallini & Scotchmer*, supra* note 38, at 54 (making and then immediately critiquing this critique).

[57] Hemel & Ouellette, *supra* note 9, at 334 (quoting Rachel Glennerster, Michael Kremer & Heidi Williams, *Creating Markets for Vaccines*, 1 INNOVATIONS: TECH., GOVERNANCE, GLOBALIZATION 67, 71 (2006)). Of course, the ability of researchers to later pursue patents on their innovations results in blending the incentive features of grants and patents.

[58] *See* Hemel & Ouellette, *supra* note 9, at 334–37. As Hemel and Ouellette note, the case of tax incentives is slightly more complicated; they create approximately ex ante incentives

does not succeed in the research, she gets nothing—patents typically provide a route to profit only if a successful product is created, prizes go only to the victor, and R&D tax incentives are usually meaningless without underlying profits. Thus, she faces incentives to conduct her work efficiently, effectively, and successfully to recoup her own expended funds. Grants, in this view, provide few incentives in the same vein.

C. Problematic Risk Allocation

The third, related critique involves the allocation of risk in grant-funded research efforts. Brett Frischmann argues that "when utilizing grants, the government, as investor-principal, often bears the entire downside risk of an unsuccessful project."[59] Because of the unconditionality of grants, when a grant-funded researcher fails to innovate, the funder has no way to recover the expended funds. This critique implicitly relies on a private-contracting analogy, where the government, as innovation funder, has the same sort of profit-and-loss incentives as a private party. The reality, as discussed below, is more complex.[60]

The other half of this critique is that the grantor also receives little of any upside benefit of successful innovation. If the government funds groundbreaking research that results in a blockbuster drug, the government receives none of the profit—and in fact, is instead likely to pay much of that drug's future cost because it pays for a large fraction of health-care costs.[61] Under an older, contrasting model, the federal government retained robust rights in research it funded, though it rarely exploited them.[62] This model largely ended with the enactment of the Bayh-Dole Act of 1980.[63] Under Bayh-Dole, grant recipients keep patent rights to federally funded research, with the rationale that these private actors can more effectively act to commercialize

that are available within the same year as the funding, but they require some source of stop-gap funding such as venture capital or other resources; and if a company fails or has no income, tax credits are worthless. *See id.* at 336–37. These concerns are mitigated by fully refundable tax credits, offered by some states. *See id.* at 337–38.

[59] Frischmann, *supra* note 31, at 387 (cited with approval by Sarnoff, *supra* note 17, at 1118).

[60] *See infra* Section IV.A.1.

[61] *See, e.g.*, Roin, *supra* note 8, at 1039–44 (describing government payments for drugs through health insurance systems).

[62] *See* Danielle Conway-Jones, *Research and Development Deliverables under Government Contracts, Grants, Cooperative Agreements and CRADAs: University Roles, Government Responsibilities and Contractor Rights*, 9 COMP. L. REV. & TECH. J. 181, 186–88 (2004) (describing the history of federal rights in funded research).

[63] Bayh-Dole Act, Pub. L. No. 96-517, 94 Stat. 3015 (1980) (codified as amended in various sections of 35 U.S.C.).

the nascent technology.[64] A vast literature considers the benefits of this move.[65] Notwithstanding whether this transfer of rights to private parties was necessary or beneficial on net, the fact remains that because the government does not retain rights to funded inventions, it lacks the ability to capture the upside of those inventions and often must pay to access them.[66]

This complaint about government inability to capture the upside of grant-funded research appears most forcefully in the public health literature, where scholars decry the lack of access to the products of government-funded research.[67] In the innovation literature, on the contrary, the cost of reduced access is often classified as a necessary evil to drive the commercialization effort.[68]

III. GRANTS IN PRACTICE (AT THE NATIONAL INSTITUTES OF HEALTH)

This Part describes how grants really work. It begins with a basic overview of the grants ecosystem. It then turns to the NIH, and describes in

[64] 35 U.S.C. § 202 (2018); *see* Conway-Jones, *supra* note 62, at 188–92 (giving a history of technology transfer legislation and executive actions). The Bayh-Dole Act addressed only universities and nonprofits. The Stevenson-Wydler Technology Innovation Act of 1980, Pub. L. No. 96480, 94 Stat. 2311 (codified at 15 U.S.C. § 1701), in a parallel structure, enabled government researchers to retain title to patents. And Executive Order 12618 extended the Bayh-Dole Act to for-profit corporations.

[65] For a few places to start, see, for example, Rebecca S. Eisenberg, *Public Research and Private Development: Patents and Technology Transfer in Government-Sponsored Research*, 82 VA. L. REV. 1663 (1996); Arti K. Rai, *Regulating Scientific Research: Intellectual Property Rights and the Norms of Science*, 94 NW. U. L. REV. 77 (1999); F. Scott Kieff, *Facilitating Scientific Research: Intellectual Property Rights and the Norms of Science—A Response to Rai and Eisenberg*, 95 NW. U. L. REV. 691 (2001); Arti K. Rai & Rebecca S. Eisenberg, *Bayh-Dole Reform and the Progress of Biomedicine*, 66 L. & CONTEMP. PROBS. 289 (2003); Emily Michiko Morris, *The Many Faces of Bayh-Dole*, 54 DUQ. L. REV. 81 (2016); DAVID C. MOWERY ET AL., IVORY TOWER AND INDUSTRIAL INNOVATION: UNIVERSITY-INDUSTRY TECHNOLOGY TRANSFER BEFORE AND AFTER THE BAYH-DOLE ACT (2004).

[66] Under § 202(c)(4) of the Bayh-Dole Act, the federal funding agency shall receive a worldwide, nonexclusive, nontransferable, irrevocable, fully paid-up license to practice the invention on behalf of the United States (or have the invention practiced). However, the Bayh-Dole Act covers only federally funded research and may not cover other patented inventions necessary to practice the innovation.

[67] *See, e.g.*, Amy Kapczynski & Aaron S. Kesselheim, *'Government Patent Use': A Legal Approach to Reducing Drug Spending*, 35 HEALTH AFF. 791 (2016) (describing the problem and proposing the use of 28 U.S.C. § 1498 (2018) to help address the concern).

[68] *See* Benjamin N. Roin, *Unpatentable Drugs and the Standard of Patentability*, 87 TEX. L. REV. 503, 507–15 (2009) (describing the rationale for patents to allow firms to recover the high costs of drug discovery); *but see, e.g.*, Glennerster et al., *supra* note 57, at 68–70, 77 (describing the desirability of minimizing deadweight loss from drug patents).

considerable detail the NIH grant-funding process, organized around the three critiques presented in Part I.

A. An Overview of Grants

External grants funded by the NIH are the focus of this Article, but some initial context is useful. The NIH is not the only funder of grants in the federal government, the federal government is not the only funder of grants, and grants are not the only way the federal government invests directly in research.

How do grants work, at a basic level? Typically, the sponsoring agency solicits applications for funding (at the NIH, frequently "requests for applications," or RFAs) at a particular level of generality, which can range from almost totally open calls for worthy research to very specific calls for proposals to address a particular issue.[69] Prospective grantees submit applications, which typically include information about their qualifications, the research they propose to undertake (often including preliminary data), and how much they expect it to cost—that is, how they expect to spend the grant funds. The grantor decides through some mechanism—much more on this later—which of the applications, if any, to fund, and then disburses the money either fully prospectively, in tranches, or as reimbursements once research expenses are incurred.[70] Often, grants come with obligations, which can range from acknowledging the funder to committing to make any resulting knowledge publicly available.[71]

Grants are not the only way the government directly funds innovation.[72] The government may also directly conduct intramural research by employing scientists at, for instance, National Laboratories or laboratories at the NIH or the Centers for Disease Control and Prevention.[73] If the federal government relies instead on non-governmental researchers, it uses grants when it wishes to fund research but does not "acquire . . . property or services for the direct benefit or use of the United States Government" and "substantial involvement" of the federal agency is not expected.[74] If the government will

[69] *See infra* Section III.B.1.

[70] *See* NIH, NIH Grants Policy Statement IIA-59 (2016) [hereinafter NIH Grants Policy Statement].

[71] *See infra* Section III.B.2 (describing disclosure requirements).

[72] Indirectly, the government funds innovation through several mechanisms already mentioned, including R&D tax credits and the enforcement of patent and trade secrecy laws (which fund research through ex post "shadow taxes" on users of the patented or secret technology). *See* Hemel & Ouellette, *supra* note 9, at 320–26.

[73] *See* Sarnoff, *supra* note 17, at 1132–36 (describing the role of government agencies in promoting research and development).

[74] 31 U.S.C. § 6304 (2018).

acquire goods or services, it uses the procurement system—a $440-billion-annual-spending behemoth[75]—instead.[76] If the federal agency expects to be substantially involved, such as in collaborations between National Laboratories and private industry, the agency uses Collaborative Research and Development Agreements (CRADAs) to direct the collaboration.[77] The federal government may also offer prizes, though these remain rare and limited.[78] While each of these different forms of direct government subsidy is substantial and important,[79] this Article focuses on federal extramural grants: the distribution of funding to innovators outside the government's walls without the expectation of government involvement or government receipt of goods or services. Such grants are especially important to university researchers.[80]

Although federal agencies are the dominant grant funders today, this was not always the case and they are not the only source of grant funding. Governments at any level, including federal, state, and local, may fund research grants.[81] Private not-for-profit organizations may also fund research grants.[82] Grants may be funded internally by universities or other research institutes out of their own funds.[83] Finally, grants may be funded by private industry, a funding source that has received increasing attention though it remains comparatively small.[84] International grant funding is similarly diverse, though

[75] NAT'L CONTRACT MGMT ASS'N, ANNUAL REVIEW OF GOVERNMENT CONTRACTING 2 (2016).

[76] *See* 31 U.S.C. § 6303 (2018); Conway-Jones, *supra* note 62, at 192-97 (detailing the rights and responsibilities of government and contractors in procurement agreements).

[77] 31 U.S.C. § 6305 (2018).

[78] *See* Hemel & Ouellette, *supra* note 9, at 317–18.

[79] *See, e.g.*, Sarnoff, *Likely Mismatch*, *supra* note 31, at 375–80 (comparing several direct sources of government funding, focusing on direct funding over market regulation like patent law).

[80] Barry Bozeman & Monica Gaughan, *Impact of Grants and Contracts on Academic Researchers' Interactions with Industry*, 36 RESEARCH POL'Y 694, 694 (2007).

[81] *See, e.g.*, *All CIRM Grants*, CAL. INST. REGENERATIVE MED., https://www.cirm.ca.gov/grants [https://perma.cc/BC2F-R6ZG] (last visited March 10, 2019) (listing grants awarded by California's state-funded stem-cell research agency). For a description of how state and local governments provide innovation financing more generally, see Hrdy, *Patent Nationally*, *supra* note 26, at 1363–75.

[82] *See* LILY E. KAY, THE MOLECULAR VISION OF LIFE (1993), *passim* (describing the support provided by the Rockefeller Foundation for the California Institute of Technology and its development of the field of molecular biology).

[83] *See, e.g.*, *MCubed*, UNIV. MICH., http://mcubed.umich.edu/ [https://perma.cc/PR8R-ZDUU] (last visited March 10, 2019) (describing the university-funded MCubed grant program for intramural research).

[84] Bozeman & Gaughan, *supra* note 80, at 694 ("[A]t no time during the history of the modern U.S. research university . . . has industry provided as much support for *university*

the relative balance between different governmental levels, not-for-profit, and for-profit funding may vary between countries.[85]

In the United States, federal research grants have grown tremendously in the last half-century.[86] In the first half of the twentieth century, private foundations provided most extramural funding; the Rockefeller Foundation, for instance, was mostly responsible for the early growth of molecular biology as a field.[87] After World War II, the federal science budget grew tremendously, and the government displaced private foundations to become the dominant funder of research.[88] Today, while the private sector spends more on research than the federal government does, it spends mostly within its own walls; the federal government remains the dominant source of extramural scientific grant funding, especially for basic research.[89]

Within the federal government, many agencies fund research through grants, including the Department of Defense, the Environmental Protection Agency, and the Department of Energy.[90] Two agencies especially focus on funding basic research: the National Science Foundation (NSF) and the NIH. The NSF funds research across many scientific fields, including substantial amounts of basic biological research.[91] But the largest funder of grant-based

research as any of the top five government funding agencies."). Private R&D funding as a whole is large, but mostly intramural. *See id.* (noting that industry is the leading source of R&D funding nationally). Nonetheless, industry grants have been perceived as having outsized importance relative to their size. *See id.* at 695; *see also* Mats Benner & Ulf Sandstrom, *Institutionalizing the Triple Helix: Research Funding and Norms in the Academic System*, 29 RES. POL'Y 291, 293 (2000) (noting how industry funding can change research trajectories).

[85] An overview of the international grant system is beyond the scope of this Article. For a few useful resources, see, e.g., Christoph Grimpe, *Extramural Research Grants and Scientists' Funding Strategies: Beggars Cannot be Choosers?*, 41 RES. POL'Y 1448, 1450 (2012) (giving an overview of the European and German grant systems); SUSAN WRIGHT, MOLECULAR POLITICS 32–36, 60–63 (1994) (giving a history of the United Kingdom's grant-funding system in the twentieth century).

[86] WRIGHT, *supra* note 85, at 21.

[87] *Id.*; *see also* KAY, *supra* note 82, *passim*.

[88] WRIGHT, *supra* note 85, at 21.

[89] *See* Mike Henry, *US R&D Spending at All-Time High, Federal Share Reaches Record Low*, AM. INST. PHYSICS (Nov. 8, 2016), https://www.aip.org/fyi/2016/us-rd-spending-all-time-high-federal-share-reaches-record-low [https://perma.cc/VJG5-GRLA] (noting that private spending reached 69% of total R&D while federal spending dropped to 23%, but also noting that the federal government remains the top funder of basic research).

[90] *See Grant-Making Agencies*, GRANTS.GOV, https://www.grants.gov/web/grants/learn-grants/grant-making-agencies.html [https://perma.cc/S8MZ-DN3N] (last visited March 10, 2019).

[91] Richard Freeman & John Van Reenen, *What If Congress Doubled R&D Spending on the Physical Sciences?*, 9 INNOVATION POL'Y & ECON. 1, 6 (2009); Thomas O. McGarity, *Peer Review in Awarding Federal Grants in the Arts and Sciences*, 9 HIGH TECH. L.J. 1, 15–16.

research by far, focusing entirely on biomedical science, is the NIH, "the center of a vast research system unmatched in size and scope throughout the world."[92] The NIH comprises twenty-seven different Institutes and Centers (collectively, "Institutes"), each focused on a "specific disease area, organ system, or stage of life"; examples include the National Cancer Institute, the National Human Genome Research Institute, and the National Institute on Aging.[93] Of these, twenty-four make grant awards.[94] The NIH expends about $37.3 billion in biomedical research per year; 10% of that is spent on its own intramural research programs, and around 80% on extramural grants.[95] "[I]n the market for biomedical research, NIH is the 800 pound gorilla."[96]

B. TESTING THE THREE CRITIQUES AT THE NIH

Part II introduced three critiques of the grant system: they rely on bureaucratic decision-making; they are largely unaccountable due to ex ante funding; and they poorly allocate risk by giving the grantor most of the downside risk and little of the upside. These critiques largely fail to reflect the reality of the modern grant system, at least as practiced at the NIH. Uninformed bureaucrats do not make the principal funding decisions, which are instead effectively made by panels of well-informed peer scientists. Funding is only ex ante and (mostly) unaccountable for single grants, but researchers are repeat players and depend on the *next* grant as well, creating accountability.[97] And rather than misallocating downside risk entirely to the NIH and the upside entirely to the researcher, the NIH actually sees much more upside benefit—and researchers more downside cost.

1. *Bureaucratic Decision-Making*

How are grant decisions made at the NIH?[98] In brief: the NIH seeks grant applications, peer reviewers evaluate and compare the grant applications

[92] WRIGHT, *supra* note 85, at 26.

[93] For a full list of the twenty-seven institutes and centers, see *List of NIH Institutes, Centers, and Offices*, NIH, https://www.nih.gov/institutes-nih/list-nih-institutes-centers-offices [https://perma.cc/5TMP-2LTN] (last visited Mar. 11, 2019) [hereinafter *List of NIH Institutes*].

[94] *Understanding the NIH: Finding the Right Fit for Your Research*, NIH, https://grants.nih.gov/grants/understanding-nih.htm [https://perma.cc/W5W6-ZGES] (last visited Mar. 11, 2019).

[95] NIH, *Budget*, *supra* note 4.

[96] Freeman & Van Reenan, *supra* note 91, at 19.

[97] As mentioned above, grants do not act in isolation; researchers may also be able to patent useful inventions, which provides an additional incentive. However, this Article focuses on incentives internal to the grant system.

[98] *See generally Grants Process Overview*, NIH, https://grants.nih.gov/grants/grants_process.htm [https://perma.cc/2VXH-ZYQW] (last visited Mar. 11, 2019).

submitted in response, and the NIH makes final funding decisions. In both the seeking of grant applications (that is, deciding what areas of innovation to fund) and the process of peer review (that is, deciding which innovators and projects specifically to fund), the NIH funding process belies the critique that grant-funding decisions are made by bureaucrats lacking relevant knowledge. This is especially true for the broad, open R01 research project grant program.[99] As Richard Freeman and John Van Reenen put it:

> At the heart of the American biomedical science enterprise are the R01 grants that the NIH gives to fund individual scientists and their teams of postdoctorate employees and graduate students. The system of funding individual researchers on the basis of unsolicited applications for research support comes close enough to economists' views of how a decentralized market mechanism operates to suggest that this ought to be an efficient way to conduct research compared, say, to some central planner mandating research topics. The individual researchers choose the most promising line of research on the basis of "local knowledge" of their special field. They submit proposals to funding agencies, where panels of experts—"study sections" in the NIH world—give independent peer review, ranking proposals in accordance with criteria set out by funding agencies and their perceived quality. Finally, the agency funds as many proposals with high rankings that it can within its budget constraints.[100]

This Section explores the grant-funding process.

a) Seeking Grant Applications

The first step of innovation funding is deciding what areas of innovation to fund. Some innovation incentives, like prizes, typically require that the target be fully identified beforehand. Others, like patents, require no ex ante identification by any administrator; private firms decide what opportunities to pursue. Grants might resemble prizes, in that the government identifies beforehand what it would like to fund. As we shall see, this is only partially true; at the NIH, some grant funding ("solicited" applications) looks like broadly-defined prizes, with innovation targets identified up front; other funding ("unsolicited" projects) resembles patents, in that the agency is open to a very wide range of possible projects. In either form, the NIH announces that it will accept applications in a "Funding Opportunity Announcement"

[99] The NIH's "R" grants provide support for research projects. *See Research Grants (R)*, NIH, https://www.nimh.nih.gov/funding/grant-writing-and-application-process/research-grants-r.shtml [https://perma.cc/3T57-AH6H] (last visited Mar. 11, 2019).

[100] Freeman & Van Reenen, *supra* note 91, at 18–19.

that lays out the parameters for what sorts of grants might be funded.[101]

Unsolicited grants allow individual innovators to suggest their own projects within very broad parameters. The NIH has created a standing set of "parent announcements" that last for a number of years, with standard application dates.[102] Under the announcements, researchers can propose their own project, so long as it fits within the very broad mission of the NIH and of the funding Institute (for instance, cancer-related research to be funded by the National Cancer Institute).[103] The broadest and most well-known of these parent announcements is the R01 Research Project Grant, which "supports a discrete, specified, circumscribed project in areas representing the specific interests and competencies of the investigator(s)."[104] Other standing parent announcements exist for smaller research projects, grants for training young scientists, fellowships, and professional development grants.[105] Overall, this set of funding represents a "deliberate policy of relying on the judgment of the scientific community as a whole, through investigator-initiated proposals, to determine the scientific agenda and identify the areas in which progress is most likely."[106] Historically, around 80 to 90% of NIH grant awards are

[101] *See infra* notes 116–119 and accompanying text.

[102] *What Does NIH Look For?*, NIH, https://grants.nih.gov/grants/what-does-nih-look-for.htm [https://perma.cc/C2TU-4FKR] (last updated May 24, 2016); *Parent Announcements (For Unsolicited or Investigator-Initiated Applications)*, NIH, https://grants.nih.gov/grants/guide/parent_announcements.htm [https://perma.cc/7JSK-UWLD] (last visited Mar. 11, 2019) [hereinafter *Parent Announcements*].

[103] Proposals must fit the mission of an NIH Institute, so unsolicited grants are not a pure free-for-all. Nevertheless, the collective set of NIH Institutes covers a very wide swath of biomedical research: Institutes focus on general medical sciences, environmental health, diseases (cancer, alcohol abuse, drug abuse, allergies, infectious diseases, arthritis, musculoskeletal disease, skin disease, deafness, diabetes, digestive disease, kidney disease, mental health, neurological disorders, and stroke), minority populations, techniques (genomic research, biomedical imagining, bioengineering, nursing, clinical research, information technology, and translational science), life stages (aging, child health, and human development) and organ systems (eyes, hearts, lungs, blood, and teeth). *List of NIH Institutes*, *supra* note 93.

[104] *NIH Research Grant Program (Parent R01), Announcement No. PA-06-160*, NIH, https://grants.nih.gov/grants/guide/pa-files/PA-16-160.html [https://perma.cc/3V9M-W2AP] (last visited Mar. 11, 2019) (announcing availability of R01 grants from 20 National Institutes as well as the National Library of Medicine; the National Center for Complementary and Integrative Health; and the Office of Research Infrastructure Programs' Division of Program Coordination, Planning and Strategic Initiatives for the three years beginning in May 2016).

[105] *See Parent Announcements*, *supra* note 102 (listing parent announcements in the R (research), T (research training), K (career development), and F (fellowships) series, among others).

[106] INSTITUTE OF MED., NIH EXTRAMURAL CENTER PROGRAMS: CRITERIA FOR INITIATION AND EVALUATION 49 (Frederick J. Manning, Michael McGeary & Ronald Estabrook eds., 2004).

unsolicited.[107]

The NIH also solicits research proposals, which look a bit more like prizes—albeit very broad prizes—inasmuch as they involve greater ex ante decision-making about what areas of innovation are worth funding. Solicited proposals are intended to address areas the agency thinks worth funding for a variety of reasons, including "to support research in an understudied area of science, to take advantage of current scientific opportunities, to address a high scientific program priority, or to meet additional needs in research training and infrastructure."[108] Soliciting research often deeply engages active researchers; Institutes frequently convene groups of scientists who discuss what research is ongoing, what opportunities exist, and what the Institute should fund.[109] One scientist described such a group conducted at the National Cancer Institute as a "really intense think tank" that realized a need "to bring different disciplines together and enable them to really think differently about cancer."[110] Once the group of scientists mapped roughly what the program should look like to accomplish this scientific/innovation goal, NCI staff "went back internal," and decided how precisely to shape the program.[111] The exact contours of this process vary substantially across Institutes.[112] Even where priorities are generated by NIH employees, many of them are trained as scientists in their own right.[113]

Solicited research programs also grow from top-down priorities. Congress

[107] NIH, *Research Project Grants: New (Type 1) Awards and Percentage to Targeted Research* (1997–2017), https://report.nih.gov/NIHDatabook/Charts/Default.aspx?showm=Y&chartId=25&catId=2 [https://perma.cc/VAF3-UVBK].

[108] *What Does NIH Look For?*, NIH, https://grants.nih.gov/grants/what-does-nih-look-for.htm [https://perma.cc/3Y4G-V9H9] (last visited Mar. 11, 2019).

[109] *See* INSTITUTE OF MED. & COMM. ON THE NIH RESEARCH PRIORITY-SETTING PROCESS, SCIENTIFIC OPPORTUNITIES AND PUBLIC NEEDS: IMPROVING PRIORITY SETTING AND PUBLIC INPUT AT THE NATIONAL INSTITUTES OF HEALTH 49–51 (1998) (describing various bottom-up procedures for setting research priorities at Institutes) [hereinafter IOM, PRIORITY SETTING].

[110] Interview with Anonymous Senior Scientist (June 7, 2017) (on file with author).

[111] *Id.*

[112] *See* IOM, PRIORITY SETTING, *supra* note 109, at 51 (noting "tremendous variability" in Institutes' "systems for receiving advice, planning, and setting priorities . . . . [S]ome institutes appear to adopt plans developed by a proactive staff with the endorsement of advisory groups, whereas others follow closely the recommendations of external advisory groups").

[113] *See* Marion Zatz, *A View from the NIH Bridge: Perspectives of a Program Officer*, 22 MOLECULAR BIOLOGY CELL 2661, 2662–63 (2011) ("Like many of my colleagues at the NIH, I came to this position following a career as an independent research scientist, where I developed many skills that are essential for being a successful researcher or teacher, and for being a [program officer].").

can directly set research priorities, either generally, by deciding how much money to appropriate to a particular Institute (and, accordingly, its broad research focus), or specifically, as the 21st Century Cures Act did in supporting the Precision Medicine Initiative.[114] The President or other White House officials can also drive priorities; President Obama directly proposed the Precision Medicine Initiative, aimed at generating and collecting the health data of a million Americans for future research purposes, and 2016's Cancer Moonshot, focused on fighting cancer.[115] The Human Genome Project was similarly the subject of high-level executive focus. The Directors of Institutes or of the NIH can also shape the agency's funding priorities.[116] Even if priorities are established politically, however, groups of active researchers are still involved in determining how the top-down priority should be implemented.

Once the funder has decided what opportunities to pursue, it issues a Funding Opportunity Announcement, typically as either a "Program Announcement"[117] or "Request for Application" (RFA).[118] A Program Announcement indicates an area of interest, and an RFA formally solicits grant applications "in a well-defined scientific area to accomplish specific program objectives."[119] It describes how much funding the NIH expects to make

[114] 21st Century Cures Act, Pub. L. No. 114-255, § 1001(b)(4)(A), 130 Stat. 1033 (2016) (appropriating $1.455 billion for the Precision Medicine Initiative); *id.* at § 2011 (amending the Public Health Services Act to "encourag[e]" the Secretary of Health and Human Services "to establish and carry out . . . the 'Precision Medicine Initiative' ").

[115] *See* Jacob S. Sherkow, *Cancer's IP*, 96 N.C. L. REV. 297 (2018) (describing the Cancer Moonshot and describing the intellectual property challenges arising in the context of cancer).

[116] The NIH Director is involved in budget negotiations with Congress; Institute Directors have final say on areas of funding emphasis and can identify special areas of emphasis. *See* NIH, SETTING RESEARCH PRIORITIES AT THE NATIONAL INSTITUTES OF HEALTH 15 (1997). In addition, the Director has substantial influence over a designated funding source, the Common Fund, aimed at areas difficult for any single Institute to address on its own. *See About the NIH Common Fund*, NIH, https://commonfund.nih.gov [https://perma.cc/GX3T-GUTX] (last visited Mar. 11, 2019).

[117] A Program Announcement is "a formal statement about a new or ongoing extramural activity or program. It may serve as a reminder of continuing interest in a research area, describe modification in an activity or program, and/or invite applications for grant support." *Glossary & Acronym List: Program Announcement (PA)*, NIH, https://grants.nih.gov/grants/glossary.htm#ProgramAnnouncement(PA) [https://perma.cc/Z6XH-GBTV] (last visited Mar. 11, 2019).

[118] A Request for Application is "a formal statement that solicits grant or cooperative agreement applications in a well-defined scientific area to accomplish specific program objectives." *Glossary & Acronym List: Request for Application (RFA)*, NIH, https://grants.nih.gov/grants/glossary.htm [https://perma.cc/NG99-9QUN] (last visited Mar. 11, 2019).

[119] *Id.* An RFA can also solicit cooperative agreement applications. *Id.*

available, how many grants it expects to fund, and other logistical details.[120]

The process of seeking applications and thereby setting innovation target areas is markedly more complicated than suggested by the critique of the grant system. There is some truth to the idea that bureaucrats are making decisions; the staff of various institutes and centers are involved in setting priorities to determine what sorts of innovation may be funded, and in crafting the actual RFAs and Program Announcements that formally invite grant applications. And "the government," writ large, can influence what areas are funded: Congress can appropriate funds for particular projects (and, indeed, appropriates funds separately for each Institute, giving it a chance to prioritize the different broad missions), and the White House has been closely involved in establishing large-scale research programs.[121] Broad political controversies can also informally shape researcher behavior.[122] But this is far from the whole story. The Parent Announcements are broad, standing invitations to seek funding for whatever projects a researcher thinks worthy of funding that fits within that capacious mission of the NIH, and a majority of research or training applications submitted to the NIH fall within such investigator-initiated categories.[123] And even for the more focused Program Announcements and RFAs, practicing scientific researchers are involved in crafting the rationale for, and the shape of, solicitation for grant applications.

b) Peer Review

The second key funding issue involves individual projects: once areas of targeted innovation have been broadly identified, what specific projects should be funded, and who should undertake those projects? These two questions are tightly blended in the NIH's peer review system, the heart of the NIH's grant evaluation system. The NIH is required by law to use peer review to evaluate grants.[124] About 25,000 peer scientists review about 80,000 grant applications

[120] *See id.*

[121] *See supra* note 115, at 299–300 and accompanying text.

[122] *See, e.g.*, Joanna Kempner, *The Chilling Effect: How Do Researchers React to Controversy?*, 6 PLoS MED. 1571, 1571 (2009) (finding that among researchers whose NIH grant proposals had been criticized as wasteful in a "highly publicized political controversy," about half later removed controversial words from grants and about a quarter avoided controversial topics); Rebecca Hersher, *Climate Scientists Watch Their Words, Hoping to Stave Off Funding Cuts*, NPR (Nov. 29, 2017), https://www.npr.org/sections/thetwo-way/2017/11/29/564043596/climate-scientists-watch-their-words-hoping-to-stave-off-funding-cuts [https://perma.cc/R5ED-4MDT] (noting a sharp decrease in the phrase "climate change" in NSF grants in reaction to the Trump administration's hostility to the topic).

[123] *See* NIH GRANTS POLICY STATEMENT, *supra* note 70, at I-46.

[124] *See* 42 U.S.C. § 289a (2018).

each year in two stages:[125] "initial peer review" for "scientific and technical merit" and Advisory Council Review, which includes broader policy considerations.[126] An application must be recommended for approval by both levels to be recommended for final funding by an Institute.[127] Of the two, the initial peer review is far more important for individual grants.

Initial peer review focuses on the science alone. When researchers submit a grant application, the NIH's Center for Scientific Review checks the application for technical details and conformance with the Funding Opportunity Announcement, then assigns the application to a Scientific Review Group for initial peer review.[128]

Scientific Review Groups (Groups) are mostly made up of non-government scientists with relevant scientific and technical expertise.[129] However, each Group is led by an NIH staff scientist, known as a Scientific Review Officer, who recruits reviewers, assigns applications to reviewers for pre-meeting review, and prepares summaries of the grant's evaluation.[130] The non-federal scientist peer reviewers receive the grant applications several weeks in advance of a peer review meeting.[131] Each is assigned particular applications to pre-review, which includes writing a critique and scoring the application preliminarily.[132]

Grant applications are scored on several criteria. The most important is "overall impact" ("likelihood for the project to exert a sustained, powerful influence on the research field(s) involved").[133] Several other criteria are scored; for research project grants, these are typically:[134]

[125] NIH, NIH PEER REVIEW: GRANTS AND COOPERATIVE AGREEMENTS (2019).

[126] *Peer Review*, NIH, https://grants.nih.gov/grants/peer-review.htm [https://perma.cc/K6ZK-WAEP] (last visited arch 11, 2019) [hereinafter NIH, *Peer Review*].

[127] *See id.*

[128] *See id.*

[129] *See id.*

[130] *See id.*

[131] *See id.*

[132] *See id.* The NIH provides copious guidance to its peer reviewers, including policies on avoiding conflicts of interest, evaluating proposal significance and impact, evaluating researcher plans to share data, and evaluating the rigor and transparency of a proposal. *See generally Consolidated List of Reviewer Documents*, NIH https://grants.nih.gov/grants/peer/reviewer_guidelines.htm [https://perma.cc/E5KX-GF8S] (last visited Mar. 11, 2019).

[133] NIH, *Peer Review*, *supra* note 126.

[134] *See Definitions of Criteria and Considerations for Research Project Grant (RPG/X01/R01/R03/R21/R33/R34) Critiques*, NIH, https://grants.nih.gov/grants/peer/critiques/rpg_D.htm#rpg_01 [https://perma.cc/9FGG-2V88] (last visited March 11, 2019) [hereinafter NIH, *Definitions of Criteria*]. Additional criteria may be provided for different grant types. *Id.*

Significance: scientific basis for the project, and how it could change and improve the field;

Investigator(s): experience and suitability of the researchers for the project, including experience and training (for young investigators) and demonstrated accomplishments (for established researchers);

Innovation: novel (in the field or broadly) paradigms, interventions, approaches, etc., to "challenge and seek to shift current research or clinical practice paradigms";

Approach: "well-reasoned and appropriate" "strategy, methodology, and analyses" and design of the project;

Environment: supportive scientific environment, including institutional support.[135]

The five criteria listed above, as well as overall impact, are numerically scored.[136] Additional criteria involve protections for human subjects, diversity, animal policies, and others, but these criteria are not scored.[137]

Once the assigned peer reviewers have given initial scores, those scores (typically just the overall impact score) are used to determine which applications will be discussed at the Group meeting; applications that do not make the cut (typically the bottom half) are "not discussed" and will not be funded.[138] At the meeting, the remaining grant applications receive a final overall impact score from each non-conflicted Group reviewer; these scores are averaged to obtain a final total score, which ranges from 10 (high impact) to 90 (low impact).[139]

The second level of peer review is by the National Advisory Council or National Advisory Board (together, "Council") associated with the potentially funding Institute.[140] Each Council comprises both scientists and public representatives with an interest in the scientific subject or disease.[141] The Council does not typically review individual grants; instead, NIH staff construct a grant-funding plan based on the results of the initial peer review

[135] *Id.*

[136] *See Notice NOT-OD-09-024: Enhancing Peer Review: The NIH Announces New Scoring Procedures for Evaluation of Research Applications Received for Potential FY2010 Funding*, NIH, https://grants.nih.gov/grants/guide/notice-files/not-od-09-024.html [https://perma.cc/CT4T-8ETT] (noting changes to grant scoring system from a 1-to-5 scale with 0.1 point increments to a 1-to-9 integer scale) (last visited Mar. 11, 2019).

[137] NIH, *Peer Review*, *supra* note 126.

[138] *Id.*

[139] *Id.*

[140] *Id.*

[141] *Id.*

scores, and the Council makes recommendations for changes.[142] The Council nominally considers broader issues, including the mission of the Institute, the balance of funding between different recipients, and priorities of different research areas.[143] However, Council review, while "not perfunctory," is "highly deferential to study section recommendations."[144]

Finally, the Director of the Institute makes the actual funding decision. This decision can be delegated, and often final decisions are made by units within an Institute (such as Divisions or Programs).[145]

Despite the formal three-stage process—initial peer review for scientific merit, Council review for broader considerations, and a Director's final call—in practice, the initial peer review almost completely determines the outcome for the vast majority of grants.[146] Applications are ranked by their final overall score, and Institute staff determine, based on available funding, what score is necessary for a grant to be funded by the Institute: the "payline."[147] For instance, if the payline for a grant is thirty, grants with final overall scores of thirty or below are typically funded, and applications with scores above thirty are not funded.[148] Paylines may also be expressed as percentile scores among all submitted grants. For many Institutes, the payline is publicly announced; the National Cancer Institute, for instance, announced that for 2016 it would fund R01 grants up to the 10th percentile and R21 exploratory grants up to the 7th percentile "without additional review."[149] There is *some* flexibility around paylines—the paylines are typically different for less-established researchers,

[142] *See id.*

[143] For instance, the Council specially reviews individual grant applications where the investigator already receives over $1 million in NIH grant funding, though this review does not constitute a funding cap. *Id.*

[144] McGarity, *supra* note 91, at 10 (citing DARYL E. CHUBIN & EDWARD J. HACKETT, PEERLESS SCIENCE: PEER REVIEW AND U.S. SCIENCE POLICY 2 (1990)).

[145] *See* Brian A. Jacob & Lars Lefgren, *The Impact of Research Grant Funding on Scientific Productivity*, 95 J. PUB. ECON. 1168, 1169 (2011).

[146] *See id.* ("Generally, grants are awarded solely on the basis of priority score."); *see also NCI Funding Policy for RPG Awards FY16*, NAT'L CANCER INST., https://deainfo.nci.nih.gov/grantspolicies/FinalFundLtrArchive/finalfundltr2016.htm [https://perma.cc/AK7F-75UQ] (last visited Mar. 11, 2019) [hereinafter NCI, *2016 Funding Strategy*] ("Peer review evaluation of scientific merit will remain the primary consideration in these funding decisions, which will be made by NCI Scientific Program Leaders . . . following discussions with program staff.").

[147] *See generally* NIH GRANTS POLICY STATEMENT, *supra* note 70, at I-73 (noting that some Institutes and Centers publish their paylines).

[148] *See* Jacob & Lefgren, *supra* note 145, at 1171 ("[T]he realized cutoff in each situation depends on the level of funding for a particular institute, year, and mechanism, along with the number and quality of applications submitted.").

[149] NCI, *2016 Funding Strategy*, *supra* note 146.

for instance,[150] and final funding decisions may involve a small fraction of "out-of-order" funding based on other priorities of the particular Institute's administration.[151] But the vast majority of grants have their fates determined by the initial peer review for scientific and technical merit. This helps address concerns of cronyism and corruption because panels of peers, not officials, largely determine funding.

Overall, then, the system differs markedly from the simplified version presented in the critique of grants. Are government bureaucrats making uninformed decisions about what scientific projects get funded? Not really. It is true that staff and leaders at the NIH are involved in the process: the Center for Scientific Review processes initial applications and assigns them to review groups, Scientific Review Officers run the Scientific Review Groups in the initial peer review, NIH staff collates scores and prepares funding reports for the Councils, and NIH Directors or their delegates make the final decisions. But the key determinant of funding is initial peer review. Several scientists with expertise in the field read applications; determine how they fare on significance, investigator qualifications, innovation, approach, scientific environment, and overall impact on the field; and write up scores, critiques, and reasoning. Then those scientists meet, discuss the most promising grants, and decide their final scores—which projects are most worthy. That's mostly it. Grants are ranked, and the grants judged most worthy are funded until the funding runs out (with a bit of wiggle room).

While the process does not involve the market aggregating private information held by firms, it does involve the aggregation of relevant information. The grant applicants themselves disclose what they know of the innovation's potential value and their own capacities in the grant application. Peer reviewers see that information, have their own information about the field, and often can directly compare projects proposed by different researchers in the same field. And agency personnel can provide broader perspectives about government information. This process is a far cry from the notion of an uninformed bureaucrat simply sitting in a room and "picking

[150] *See id.* (noting that grants submitted by "early stage investigators" (discussed *infra* at Section IV.A.1) would be funded up to the 12th percentile, rather than the 10th percentile for other investigators); Jacob & Lefgren, *supra* note 145, at 1171 (noting that "there is clearly evidence of out-of-order funding. In [their] sample [of grant applications], 4% of individuals who scored above the cutoff received the grant, while 9% of those below the cutoff did not receive a grant or declined the award").

[151] *See* Jacob & Lefgren, *supra* note 145, at 1169 ("Institute directors have the discretion to fund applications out of order on the basis of their subjective judgment of application quality, or other factors such as how an application fits with the institute's mission or whether there were a large number of applications submitted on a similar topic.").

winners."

c) Concerns of Peer Involvement in Funding Decisions

Peer review of grants certainly brings its own challenges, including bias, conformity, and accurate prediction, some of which parallel problems raised with peer review of research publications.[152] First, bias is frequently raised as a concern. Grant applications are generally not anonymous, not least because the funding decision depends in part on the qualifications of the researchers seeking funding. Because peer scientists are involved in deciding which projects receive funding, their decisions could be biased by personal animosity,[153] prejudices against the personal characteristics of the researcher seeking funding,[154] competitiveness against researchers in the same field,[155] political pressure,[156] or otherwise. Studies have found varying levels of evidence for such bias.[157]

Second, peer review may create subtle pressure against innovative science: peers may prefer grant proposals that do not rock the scientific boat.[158] Thomas Kuhn, an influential sociologist of science, noted that the scientific model involves communities of experts making their own decisions about what research would progress.[159] Nicolas Rasmussen notes that leaving those decisions in the hands of top scientists can have the effect of concentrating

[152] *See generally* CHUBIN & HACKETT, *supra* note 144 (providing a review of peer review).

[153] *See* McGarity, *supra* note 91, at 5.

[154] *See, e.g.*, Erika C. Hayden, *Racial Bias Haunts NIH Grants*, 527 NATURE 286 (2015) (finding evidence of racial bias for NIH grant funding); Anna Kaatz et al., *Analysis of National Institutes of Health R01 Application Critiques, Impact, and Criteria Scores: Does the Sex of the Principal Investigator Make a Difference?*, 91 ACAD. MED. 1080 (2016) (finding little bias for R01 initial grants, but bias against women for R01 renewals).

[155] *See* McGarity, *supra* note 91, at 52–54 (noting the potential for financial or research conflicts of interest). *But see Managing Conflict of Interest in NIH Peer Review of Grants and Contracts*, NIH, https://grants.nih.gov/grants/peer/peer_coi.htm [https://perma.cc/NEN5-L447] (last visited Mar. 11, 2019) (describing NIH policies for avoiding peer reviewer conflict of interest and providing links to several relevant policies).

[156] *See* McGarity, *supra* note 91, at 7.

[157] *See* Simon Wessely, *Peer Review of Grant Applications: What Do We Know?*, 352 LANCET 301, 304 (1998) (reviewing sixty-one papers on bias in grant applications and concluding, "[t]he main charge against peer review, that of institutional or sex bias, is generally unfounded, with a few exceptions"). *But see* Hayden, *supra* note 154 (noting evidence of racial bias); Kaatz, *supra* note 154 (noting evidence of sex bias).

[158] *See, e.g.*, Joshua M. Nicholson & John P.A. Ioannidis, *Conform and Be Funded*, 492 NATURE 34 (2012); Michal Shur-Ofry, *Nonlinear Innovation*, 61 MCGILL L.J. 563, 577–78 (2016) (describing resistance among grantors to paradigm-shifting innovation).

[159] THOMAS S. KUHN, THE STRUCTURE OF SCIENTIFIC REVOLUTIONS 37 (2d ed. 1962); *see also* NICOLAS RASMUSSEN, GENE JOCKEYS: LIFE SCIENCE AND THE RISE OF BIOTECH ENTERPRISE 24 (2014) (discussing Kuhn).

scientific credit, power, and money.[160] McGarity draws out the implications of this for peer review of grant applications: "An important battleground in the war between the [new and old scientific] paradigms is the discretionary grants process. People who have spent their careers conducting research aimed at bolstering and extending the dominant paradigm are reluctant to direct resources toward research aimed at destroying it."[161] There may therefore be a preference toward more "mainstream" research proposals over those which buck convention.[162] Frischmann also notes this concern, arguing that innovation may suffer because of competitiveness of the grant system and the need for relatively "safe" proposals to ensure funding.[163]

The NIH explicitly fights back against any tendency to prioritize "safe" science; reviewers are required to numerically score a grant proposal for innovation, including the question, "Does the application challenge and seek to shift current research or clinical practice paradigms by utilizing novel theoretical concepts, approaches or methodologies, instrumentation, or interventions?"[164] In addition, grant programs can specifically prioritize boundary-crossing interdisciplinary work, as described below.[165] But the concern persists.

Third, some doubt whether peer review is accurate: is it good at sorting out good ideas and grant applications from bad ones? The answer to this seems to be a cautious and qualified "yes." Figuring out whether peer review accurately identifies projects most likely to succeed is challenging; basic research, in particular, is typically likely to fail, and paradigm-changing research is perhaps the most likely to fail, almost by definition.[166] Evidence suggests that peer review can probably discriminate sound applications from seriously flawed applications.[167] However, beyond that distinction, scholars debate whether better-scored grants are actually more productive.[168]

[160] RASMUSSEN, *supra* note 159, at 24.

[161] McGarity, *supra* note 91, at 41.

[162] *Id.* at 40; *see* Pedraza-Fariña & Bair, *supra* note 17, at 1097 (identifying this problem and describing it as an anti-innovation "research priority norm").

[163] Frischmann, *supra* note 31, at 389 n.184 (citing STAFF OF HOUSE COMM. ON SCIENCE, 105TH CONG., 1ST SESS., UNLOCKING OUR FUTURE: TOWARD A NEW NATIONAL SCIENCE POLICY 19–20 (Comm. Print 1998)).

[164] NIH, *Definitions of Criteria*, *supra* note 134.

[165] *See infra* Section IV.B.3.

[166] *See* Nelson, *supra* note 25, at 304.

[167] *See* Ferric C. Fang, Anthony Bowen & Arturo Casadevall, *NIH Peer Review Percentile Scores are Poorly Predictive of Grant Productivity*, 5 ELIFE e13323 (2016).

[168] *Compare id.* (finding little relationship between percentile score and grant productivity), *with* Danielle Li & Leila Agha, *Big Names or Big Ideas: Do Peer-Review Panels Select the Best Science Proposals?*, 348 SCIENCE 434 (2015) (finding a strong relationship between those

Overall, deep peer involvement, whether in the process of seeking applications and therefore identifying areas of innovation (broad or narrow) or in the process of choosing projects and individuals to fund, casts a substantially different light on the grant-funding process. Peer involvement in picking projects has its flaws; it might involve bias, it might suggest safe science, and it is certainly imperfect at identifying the best projects for funding. Similarly, the process of identifying areas of potential innovation, which relies both on peer involvement and on targeting by agency or other government actors with an eye toward social welfare priorities (or patronage, or pork), has its own flaws and idiosyncrasies. But of course, so does the principal alternative—a market-based system that relies on the incentives of private actors to decide what innovation is best to pursue, not based on evaluations of scientific merit or social welfare value, but on a calculus of what profits are appropriable through an imperfect intellectual property system or otherwise.[169] Grants aren't perfect; they're just different, and more interestingly different than is often assumed.

2. *Unaccountable Ex Ante Incentives*

Grants provide complex incentives for innovative effort. Several accounts critique grants as providing essentially only ex ante incentives, which may be less effective in motivating research effort because the innovator has fewer incentives to work efficiently.[170] As with grant funding, however, grant spending is more complicated. First, the NIH uses some modest tools to ensure that researchers are in fact working on what they proposed. Generally applicable anti-fraud laws also limit what researchers can do with government money, but typically apply only to behavior that significantly deviates from the purposes of the grant.[171] Second, and far more important, grants are not one-

measures).

[169] *See, e.g.*, Kapczynski & Syed, *supra* note 43, at 1907 (arguing that "patent rights have the potential to predictably and systematically distort private investment decisions over innovations by overstating the value of highly excludable information goods and understating the value of highly nonexcludable ones"); Ofer Tur-Sinai, *Technological Progress and Well-Being*, 48 LOY. U. CHI. L.J. 145, 156–59 (2016) (cataloging scholarly critiques of patents and markets as an innovation allocation mechanism); *id.* at 161–75 (arguing that even if patents and markets did well in satisfying preferences, they still do a relatively poor job of increasing well-being).

[170] *See supra* Section II.B.

[171] *See* False Claims Act, 31 U.S.C. §§ 3729–3733 (2018) (prohibiting making false claims); United States *ex rel.* Feldman v. van Gorp, 697 F.3d 78 (2d Cir. 2012) (finding fraud when an NIH-funded fellowship program at Cornell Weill Medical College deviated substantially from the grant application and continuing reports); U.S. Office of Inspector Gen., *Grant Fraud*, U.S. DEP'T HEALTH & HUM. SERVS., https://oig.hhs.gov/fraud/grant/index.asp [https://perma.cc/96XZ-TWEW] (describing grant fraud generally).

off events: researchers work as repeat players within a grant ecosystem where getting the next grant is an ongoing career imperative, and getting that next grant depends on productive outcomes from the current grant.[172]

a) Progress and Reporting Obligations

Grants do come with *some* continuing obligations that allow monitoring and control by the NIH. Rarely, grants have explicit requirements for progress that the NIH requires before additional funding is disbursed. For instance, a request for applications for high-risk, high-reward HIV vaccine research grants states that each application must include explicit Go/No-Go success criteria to be evaluated by the end of the second year of the nominally four-year grant; if the Go criteria are not met, the grant winds down with substantially decreased funding.[173] The center grants supporting the Human Genome Project also had robust accountability and control mechanisms to help drive a broad, expensive, collaborative enterprise.[174] But these mechanisms are unusual; most NIH grants include little more than reporting requirements.[175]

The NIH usually requires that grant recipients submit financial and progress reports at least annually.[176] Recipients must also disclose whether any potentially patentable inventions were made in the funded project, both under the Bayh-Dole Act and independently under NIH policy.[177] Grant recipients

[172] In addition, of course, the grant system does not exist in isolation; researchers who produce patentable inventions can patent them and receive some of the resulting royalties. *See supra* notes 63–66 (discussing the Bayh-Dole Act). I view this incentive as one created by the patent system, however, and not as one internal to the grant system.

[173] *Request for Application PAR-16-171: Innovation for HIV Vaccine Discovery (R01)*, NAT'L INST. ALLERGY & INFECTIOUS DISEASES, https://grants.nih.gov/grants/guide/pa-files/PAR-16-171.html [https://perma.cc/TXC6-MK5M] (last visited Mar. 11, 2019) [hereinafter, NIAID, *HIV RFA*].

[174] *See* STEPHEN HILGARTNER, REORDERING LIFE: KNOWLEDGE AND CONTROL IN THE GENOMICS REVOLUTION 96–98 (2017) (noting that genome sequencing centers would be subject to annual progress reports, frequent scientific reviews, meetings with NIH Center Directors, and rigorous evaluations on which future funding would be contingent); *id.* at 98–104 (detailing scientific evaluation strategies).

[175] *See* NIAID, *HIV RFA*, *supra* note 173 ("[A]pplications should be very different from conventional investigator-initiated R01 applications . . . . Applications that do not include Go/No-Go decision criterion/criteria will be considered incomplete and will not be reviewed.").

[176] *See* NIH GRANTS POLICY STATEMENT, *supra* note 70, at IIA-135. For many rewards, including R01 grants, financial reports need only be submitted at the end of the full grant period. *Id.* at IIA-125–26.

[177] *See id.* at IIA-130; *see also* 35 U.S.C. § 202 (2018). *But see* Arti K. Rai & Bhaven N. Sampat, *Accountability in Patenting of Federally Funded Research*, 30 NAT. BIOTECHNOLOGY 953 (2012) (noting that many Bayh-Dole reporting mandates go unfollowed).

are also subject to audit.[178] Failure to follow reporting requirements, or failure to comply with other terms of the grant, can theoretically result in disallowing costs, withholding future grant awards, suspending the grant, or even terminating the grant.[179] At least in part, these reporting requirements should encourage grant recipients to work toward the goals of the grant, in contrast to a purely ex ante award with no oversight or reporting mechanisms at all.

b) Repeat Players

The most important reporting of grant progress comes not in response to the current grant but in applying for the next grant. Grants terms are measured in years; researcher careers are measured in decades (or, at least, most researchers hope so). Failure to get subsequent grants can result in the downsizing of a lab or the end of a career, making researchers repeat players.[180] As Gallini and Scotchmer noted, the "moral hazard" of non-contingent ex ante funding for a *single* grant "is overcome because future grants are contingent on previous success."[181] They argue that in practice, grants "operate much like prizes, with the wrinkle that a researcher must convince the sponsor in advance that his output might be worthy of a prize. For this purpose, his reputation might suffice, and in some cases, much of the research has already been completed."[182]

NIH grant-funding policy follows this pattern. The NIH scores grant applications on five main criteria, including "Investigator(s)" (the scientist running the project). "If [non-established], do they have appropriate experience and training? If established, have they demonstrated an ongoing record of accomplishments that have advanced their field(s)?"[183] In addition, many NIH grant types effectively require substantial preliminary data, which

[178] *See* NIH GRANTS POLICY STATEMENT, *supra* note 70, at IIA-143–46.

[179] *See id.* at IIA-135. Grant termination is rare, though NIH does not track such occurrences. *See* Jef Akst, *Wanted: Records of Revoked Grants*, SCIENTIST (Jan. 20, 2010), https://www.the-scientist.com/the-nutshell/wanted-records-of-revoked-grants-43553 [https://perma.cc/6EB7-GM4X]; *cf.* Jef Akst, *3 Calif Stem Cell Grants Revoked*, SCIENTIST (Nov. 3, 2009), https://www.the-scientist.com/the-nutshell/3-calif-stem-cell-grants-revoked-43763 [https://perma.cc/3TPS-QBGC] (noting the revocation of three grants by the California Institute of Regenerative Medicine for insufficient progress).

[180] *See* Adam Ruben, *Another Tenure-Track Scientist Bites the Dust*, SCIENCE (Jul. 19, 2017), http://www.sciencemag.org/careers/2017/07/another-tenure-track-scientist-bites-dust [https://perma.cc/777E-4M9V] (giving an example of how failure to get a grant can end a career).

[181] Gallini & Scotchmer, *supra* note 38, at 54.

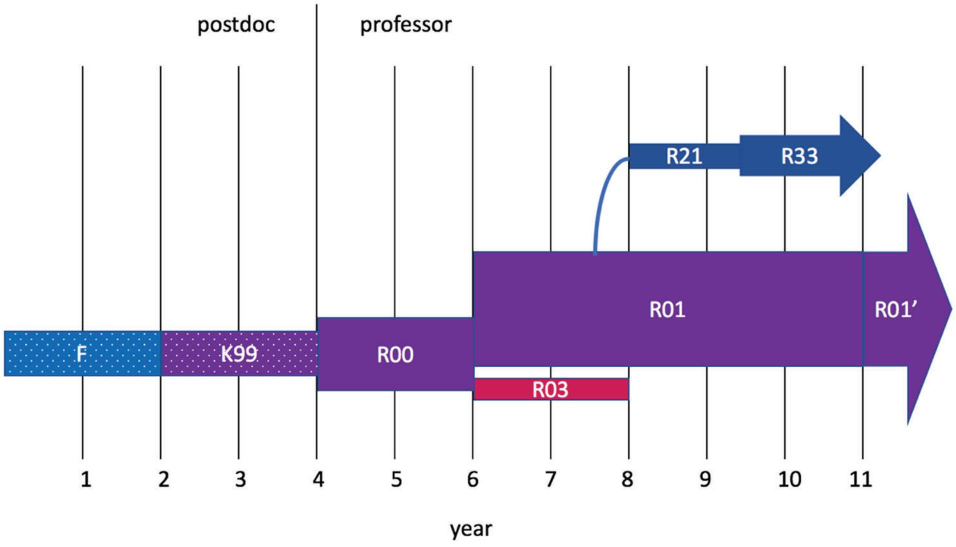
[182] *Id.*; *see also* Hanson, *supra* note 52, at 5 ("[C]ompetitive grants, which fund much of today's best basic research, can be viewed as a small prize for thinking up a promising topic, coupled with a larger but still moderate grant for working on that topic.").

[183] NIH, *Definitions of Criteria*, *supra* note 134.

serves to demonstrate (a) the project's feasibility; (b) the researcher's training and ability to generate data; and (c) the researcher's willingness to spend resources on the project even before this grant is funded.[184] This last point is in some tension with the idea that grants help free researchers from capital constraints,[185] but reinforces the serial nature of grant funding. Productivity under one grant—experiments conducted, expertise acquired, data generated, and papers published—is relevant to the NIH in deciding whether to fund the next grant, whether a competitive extension of the same project, a new grant for a related project, or an entirely different project led by the same experienced, productive researcher.

Figure 1 illustrates this pattern. It schematically shows the grants that might be received by a (rather successful) hypothetical researcher; we'll call her Jenn.

Figure 1: Schematic of serial and parallel grants



[184] Many but not all grant types require preliminary data; for instance, R01 grants require fairly substantial preliminary data, but grants focused on small studies or phased innovation (R00, R21, and R21/33 grants) need not include preliminary data, particularly if the projects are exploratory or pilot studies. *See id.*

[185] *See* Hemel & Ouellette, *supra* note 2, at 308. Without initial resources, securing preliminary results to obtain grant funding can be hard to do. The repeat nature of grants, discussed in the next Section, somewhat obviates this concern, with two caveats. First, it does not apply to initial entry to the grant system, and therefore may penalize new innovators who lack the resources to generate preliminary data on their first projects (especially if, unlike the example to follow, they do not follow a research-intensive path into becoming an innovator). Second, it may shape the direction of research, because preliminary results may not support future projects that are very far afield from the earlier work.

Initially, Jenn is supported by an F32 postdoctoral fellowship, which supports postdoctoral research and training. Jenn is working on Project Blue, under the direction of the head of her lab, Durona (the fact that Jenn is not the principal investigator is indicated by the stippling in the figure); Durona also certainly has her own funding, which supports Project Blue. Jenn uses the data acquired from that work to propose a related project, Project Purple; she applies for a K99/R00 Career Development Award, designed to help her transition into the role of an independent researcher. Getting this type of two-tiered award is contingent on Jenn's baseline qualifications, but also on how well she has done in her earlier work. It is therefore extremely challenging to get a K99/R00 grant without a record of peer-reviewed scientific publications (as well as a solid research plan and the other requirements for a grant).[186] Jenn gets the combination grant, and for two years she is funded by the K99 as a postdoctoral fellow in Durona's lab, still working under her mentorship (as the K99 requires). Then, contingent on Jenn's appointment as an independent, full time faculty member, she receives R00 funding to continue work on a broadened Project Purple in her own lab.

Two years later, Project Purple has borne fruit; the main project has developed, resulting in publications, more data, and more possibilities, and Jenn is ready to expand the project substantially. She applies for and receives an R01 Research Project Grant to continue and expand the main thrust of Project Purple: five years of substantial funding, enough to support a doctoral student and a postdoctoral fellow. But again, getting the R01 depends in large part on Jenn's research productivity while supported by the R00. Five years after getting the R01 for Project Purple, it expires; Jenn applies for a renewal (R01'), which is subject to the normal competitive grant process. For the continuation of Project Purple, Jenn's lab, and Jenn's own scientific career, productive work under each grant is essential. This is not to say that *success* is essential; the NIH knows that innovative research often fails. But future grants depend on actually doing the work.

Cross-grant contingency is not only serial but also parallel: many researchers work on multiple grants simultaneously. In an academic lab, the principal investigator who heads the lab may have working with him multiple doctoral students, multiple postdoctoral fellows, and perhaps a few technicians, working on different projects and supported by different grants—

[186] The overall success rate for 2017 K99 applications was 23.4%, but that already excludes all the candidates who did not apply because their credentials were insufficient. *See* NIH, CAREER DEVELOPMENT AWARDS: APPLICATIONS, AWARDS, SUCCESS RATES, AND FUNDING, BY INSTITUTE/CENTER AND ACTIVITY CODE (2018).

all of which partially support the principal investigator herself. Typically, these grants will be staggered in time. Even if productive results from one grant are not directly prerequisite for a staggered grant on a different project, outcomes such as papers, awards, expertise, and prizes all matter in determining whether the investigator is likely to succeed in the parallel project, and therefore whether the funder should approve that other application.

Figure 1 also shows this dynamic. At the end of Jenn's R00, she has developed another interesting line of research, and applies for an R03 Small Grant to pursue it. Unfortunately, Project Red doesn't pan out and two years later the funding runs out. Meanwhile, Project Purple continues; a couple of years later, it suggests another line of inquiry, and Jenn applies for and receives an R21 Exploratory/Developmental Grant for Project Navy (that grant requires no preliminary data, but she uses some evidence from Project Purple to support the application anyway). After two years, she has enough data from the R21 on Project Navy to get the R33 Exploratory/Developmental Grant Phase II. For each of these parallel applications, Jenn doesn't have the same sort of robust earlier data that needs to underline the serial line of Project Purple grants above. But when the Scientific Review Group conducts its initial peer review of her application,[187] it will see what she published in the course of her Project Purple work, expertise she has acquired, the experience of any postdoctoral fellows she has hired to do work, and similar progress markers. They all matter for her success as a researcher, and they all matter to peer reviewers for other grants.

In sum, while the ex ante nature of any *one* grant largely follows the critique that grants have limited ability to drive post-award researcher effort, no single grant paints the whole picture. Instead, researchers are repeat players in a system where multiple grants matter, both in parallel and serially, on the same or related projects. In this broader context, the success or productivity of work under a particular grant has far-reaching consequences on future funding, both for the researcher and for others working in her lab.

3. Problematic Risk Allocation

The third critique suggests that grants poorly allocate downside and upside risk between the funder and the recipient; a more comprehensive understanding of what benefits and costs are relevant to the NIH suggests that this critique, too, is incomplete. Of course, much of the point of grants is that the government *explicitly* does not benefit directly from successful projects (if

[187] *See supra* Section II.B.1.

it did, these instruments would be procurement contracts instead of grants).[188] Instead, grants have long been considered a way to create public goods from which the government does not directly benefit. Nevertheless, some have raised the concern that the allocation of downside and upside risk is problematic.[189] Poorly allocated risk can raise problems in both directions. If downside risk (that is, the risk of failure) is allocated entirely to the government, the researcher has decreased incentives to avoid failure. And if upside risk is allocated entirely to the researcher, the government may not reap much from its spending. Taken together, these two sides of risk allocation could also encourage researchers to pursue overly risky plans, since they capture most of the benefit of success but face little of the cost of failure. A richer conception of the grant system and the NIH's general mission reduces all three concerns.

Downside risk invites the most straightforward rebuttal. The government does not bear downside risk alone. Researchers also face downside risks from project failures. While the NIH does not require success from its funded projects—science is risky, and innovative science more so—nevertheless it is easier to generate data, and especially to publish in prestigious peer-reviewed journals, if research achieves its stated goals. This bias in favor of positive results has its own powerful negative consequences for science,[190] but it does keep some of the risk of failure squarely on the researcher. Failing to receive or renew grant funding results in a range of consequences that can hit a researcher hard, including shame among peers, the inability to hire (or the need to fire) subordinates, denial of tenure or promotion, and the end of a lab and a career.[191]

The question of upside risk allocation shifts substantially when taking into account the NIH's mission.[192] Consider an expensive NIH investment in research that leads to the development of a new drug. In all likelihood, a drug company licenses the exclusive rights to that drug, takes it to market, and reaps

[188] *See supra* note 76 and accompanying text.

[189] *See supra* Section II.A.3.

[190] An exploration of the negative repercussions of the publication bias for favorable results is fascinating but outside the scope of this Article. For an introduction to the area, see Michal Shur-Ofry, *Access-to-Error*, 34 CARDOZO ARTS & ENT. L.J. 357 (2016); John P.A. Ioannidis, *Why Most Published Research Findings Are False*, 2 PLoS MED. 696 (2005); *see also* Jacob S. Sherkow, *Patent Law's Reproducibility Paradox*, 66 DUKE L.J. 845, 852–65 (2017) (discussing the related problem of irreproducibility in science).

[191] *See* Ruben, *supra* note 180.

[192] *See Mission and Goals*, NIH, https://www.nih.gov/about-nih/what-we-do/mission-goals [https://perma.cc/8WD4-9Q5D] (last visited Mar. 11, 2019) [hereinafter NIH, *Mission and Goals*].

billions in profit while the NIH and the government see no profits; in fact, the latter pays billions to the drug company through public health insurance.[193] This dynamic is subject to a powerful critique—why doesn't the government benefit from its grant funding?[194] Simply put, it does.[195]

There are several upsides to research that the government is well positioned to capture. The simplest is that research may solve a public problem; a new vaccine will keep people from getting sick, and the government may benefit both monetarily (paying less to take care of sick people) and in its role as representative of the public (which benefits by being healthier).

A second, well-recognized benefit is that research generates information that is a public good with substantial externalities; this is perhaps the strongest justification for grants generally.[196] This is true both for basic research, the value of which is very hard to capture but which enables other innovation, and for applied research, which creates the same sort of knowledge spillovers.[197] Generating this knowledge accords with the NIH's mission, which includes "expand[ing] the knowledge base in medical and associated sciences."[198] More broadly, the government in its role as social welfare coordinator and social

[193] *See, e.g.*, U.S. GEN. ACCOUNTING OFFICE, TECHNOLOGY TRANSFER: NIH-PRIVATE SECTOR PARTNERSHIP IN THE DEVELOPMENT IN TAXOL 13 (2003) ("NIH Invested Heavily in Taxol-Related Research, but Federal Financial Benefits Have Been Limited."). While Taxol related from a cooperative research and development agreement (CRADA) rather than a grant, the argument is essentially parallel, and recurs today. *See, e.g.*, Matt Richtel & Andrew Pollack, *Harnessing the U.S. Taxpayer to Fight Cancer and Make Profits*, N.Y. TIMES (Dec. 19, 2016), https://www.nytimes.com/2016/12/19/health/harnessing-the-us-taxpayer-to-fight-cancer-and-make-profits.html?_r=1 [https://perma.cc/5349-B9C7] (asking, about government investment in CAR-T immunotherapy for cancer, "Are taxpayers getting a good deal?").

[194] *See* Mariana Mazzucato, *How Taxpayers Prop up Big Pharma, and How to Cap That*, L.A. TIMES (Oct. 27, 2015), http://www.latimes.com/opinion/op-ed/la-oe-1027-mazzucato-big-pharma-prices-20151027-story.html [https://perma.cc/N7FS-6XU5]; Gerard Anderson, *Big Pharma Should Support the NIH*, BALT. SUN (Apr. 17, 2015), http://www.baltimoresun.com/news/opinion/oped/bs-ed-medical-innovations-act-20150417-story.html [https://perma.cc/6PRP-8WNH].

[195] Issues of drug pricing are complex and far outside the scope of this piece. For an overview, see generally Ari B. Friedman & Janet Weiner, *What's the Story with Drug Prices?*, PENN LDI (May 30, 2016), https://ldi.upenn.edu/healthpolicysense/what%E2%80%99s-story-drug-prices [https://perma.cc/7FCT-7GRU].

[196] *See* Nelson, *supra* note 25, at 302–04.

[197] *See* Frischmann, *supra* note 31, at 389 ("The uncontrollable risks are borne by the government and are, in a sense, considered small because spillovers are welcome."); *cf.* Danielle Li, Pierre Azoulay & Bhaven N. Sampat, *The Applied Value of Public Investments in Biomedical Research*, 356 SCIENCE 78, 78–80 (2017) (finding that around 10% of NIH grants are directly cited by patents and 30% are cited in publications that are themselves cited in patents; for patents on approved drugs, the rates are around 1% and 5%, respectively).

[198] NIH, *Mission and Goals*, *supra* note 192.

representative realizes the benefits of those knowledge spillovers.

The public good of knowledge spillovers, however, is broader than that created by successful research. Negative information—what doesn't work, what paths are unproductive, and the like—is also useful information to both the government as a whole and to the NIH in particular. Among other things, it can help future grantees avoid fruitless research paths. Negative information can also be difficult for private firms to capture.[199]

Finally, and this upside is less often acknowledged, a substantial goal for the NIH is to build human and institutional capital in science. The NIH states that one of its four goals is "to develop, maintain, and renew scientific human and physical resources that will ensure the Nation's capability to prevent disease."[200] As a matter of both national and NIH policy, we want more trained scientists around. Their knowledge and expertise helps drive innovation across many fields. It is a positive outcome when the NIH funds, trains, and develops scientists, even if research projects fail to produce immediately valuable findings. Jacob and Lefgren find empirical evidence of successful grant-funded development: receipt of a postdoctoral fellowship (NIH's F32 grant) increases the chance of a young scientist becoming a successful researcher by almost a quarter.[201]

These two realities of downside and upside risk—that researchers do suffer from failed projects and that even risky projects can generate negative knowledge and human capital—address the concern that risk allocation will push researchers toward overly risky research projects. However, even if risk allocation *does* push researchers toward riskier projects, such an effect may be justified for two reasons. First, a risk-allocation-based push toward riskier research may counterbalance the possibility that grant-funders could prefer "safer" research.[202] Second, riskier research is likely to be a less attractive target for private investment;[203] to the extent that grant funding is especially

[199] *See* Kapczynski & Syed, *supra* note 43, at 1926–28 (noting the difficulty of capturing the benefits of negative knowledge through patents). *But see* Laura Pedraza-Fariña, *Spill Your (Trade) Secrets: Knowledge Networks as Innovation Drivers*, 92 NOTRE DAME L. REV. 1561, 1597–98 (2017) (discussing firm ability to capture negative knowledge through trade secrecy).

[200] NIH, *Mission and Goals*, *supra* note 192.

[201] Brian A. Jacob & Lars Lefgren, *The Impact of NIH Postdoctoral Training Grants on Scientific Productivity*, 40 RES. POL'Y 864, 873 (2011).

[202] *See* Pierre Azoulay, Joshua S. Graff Zivin & Gustavo Manso, *Incentives and creativity: evidence from the academic life sciences*, 42 RAND J. ECON. 527, 531 (2011) (noting NIH grant funding incentives to pursue comparatively safe research). *But see* Hyunwoo Park, Jeongsik Lee & Byung-Cheol Kim, *Project selection in NIH: A natural experiment from ARRA*, 44 RES. POL'Y 1145, 1158 (2015) (finding that NIH selects and funds riskier projects than expected).

[203] Nelson, *supra* note 25, at 302–04.

appropriate where private firms are unlikely to invest, riskier research needs grant funding more.

* * *

The NIH's vast system of grant funding reflects a richer and more complex reality than is captured in common depictions and critiques of grants. Funding decisions are not made principally by bureaucrats, but rather by panels of peer scientific experts working in concert with agency staff. Researchers respond to grant incentives not in a one-off, wholly ex ante vision that provides little drive for efficiency or success, but rather in an iterative context of serial and parallel grants where researchers are repeat players and success matters in receiving the next essential grant. And grants do not allocate downside risk just to the government and upside risk just to the grantee, but rather allocate a combination of upside and downside risks to each party.

To be sure, the experience of the NIH does not demonstrate that these critiques *never* hold—just that they do not hold at the NIH. A comparative survey of different grant systems is outside the scope of this work. However, there is reason to think that these insights are relatively generalizable. Peer review is widely used to allocate grant funds.[204] Where grant awards depend in part on prior work, and where such awards are insufficient to individually support an entire career, the repeat-player nature of the grant system should create accountability mechanisms—and those two conditions are likely to hold in most contexts. Finally, in most grant systems the recipients are likely to experience some downside risk of project failure (for the same reason), and the government to experience upside benefits.

Overall, grants are a more nuanced policy instrument than these critiques reflect. The next Part describes how they can and do help promote a broad set of innovation goals.

IV. GRANTS AS INNOVATION LEVERS

Grants can do much more than is commonly recognized. In fact, they already do. The two Sections of this Part each focus on one of the two key comparative strengths of grants: creating incentives for goods whose social welfare exceeds appropriable market value and directly supporting the development of innovation enablers. The paradigmatic version of a grant, in

[204] *See, e.g.*, McGarity, *supra* note 91, at 15–37 (describing peer review systems at the National Science Foundation, the Environmental Protection Agency, and the National Endowment for the Arts); Grimpe, *supra* note 85, at 1450–51 (noting the presence of peer review in the German scientific grant system).

the NIH context, does both of these things: an R01 basic research grant creates information useful principally for later innovation, and markets value that information for less than its social-welfare value. Because this is the paradigmatic version, on which most conceptions of government grants are based and which has been the dominant version throughout the rest of this paper, I do not describe it in detail. Instead, this Section focuses on ways grants can, do, and could promote innovation in non-paradigmatic ways.

A. SOCIAL/MARKET VALUE MISMATCHES

Grants provide a useful tool to create incentives where social value exceeds appropriable market value. This comparative strength neatly inverts the lauded ability of patents and other exclusivity mechanisms to use market signals of social value. Patents, the argument goes, are useful and efficient innovation incentives because the value a firm can realize from a patented innovation increases with the social value of the innovation, as measured by the market price and demand for that innovation.[205] But of course that argument doesn't always hold. Sometimes—some very important times—the value a firm can capture through patents doesn't reflect the social value of the innovation. One such mismatch exists when market demand fails to reflect social value because of a lack of willingness or ability to pay, as with treatments for diseases of the poor. A second mismatch happens when, although market demand might match social demand, existing appropriation mechanisms do not allow firms to appropriate an innovation's value—in effect, when existing intellectual property mechanisms fail, as with medical diagnostics.[206] Two sets of requirements shape the NIH's ability to drive innovation in these areas: the Bayh-Dole Act's requirements governing patent rights in innovations funded by government grants and the NIH's data-sharing requirements.

The Bayh-Dole Act allows universities to retain rights to inventions funded by federal grant money.[207] Instead of the federal government retaining patent rights, the Bayh-Dole Act lets universities or other nonprofits patent grant-funded innovations and license the patents to private firms for development.[208] The scheme aims to promote the commercialization of inventions by private firms, though the extent to which Bayh-Dole is necessary or beneficial is the

[205] *See supra* notes 36–38 and accompanying text.

[206] *See supra* note 21; *infra* Section III.A.2. One can also describe infrastructure investment, with its positive externalities, as a good whose appropriable market value does not scale with its social value. Because grants target innovation infrastructure and other enablers in a particularly distinct way, this opportunity is discussed in the next Section.

[207] 35 U.S.C. § 202.

[208] *Id.* For-profit grant recipients were added to the scheme by executive order. Exec. Order No. 12618, 52 C.F.R. 48661 (1987).

subject of considerable debate.[209] The government retains the right to "march in" and license the invention to another licensee if it is not made "reasonably available" by the commercializing entity, and also retains a nonexclusive license to make the invention available for government purposes.[210] The march-in right, however, has never been exercised,[211] and the government's own licensing ability has long laid dormant, though recent scholarship has attempted to revive it.[212] As Ayres and Ouellete note, the Bayh-Dole regime may have the effect of using public funding to create public goods, but then creating rewards greater than needed to develop them in the private context.[213]

The NIH's data-sharing policies also shape the availability of the fruits of grant-funded research. NIH policy requires researchers to make peer-reviewed publications resulting from grant-funded research freely available to the public one year after initial publication.[214] In addition, any "unique research resources" made with NIH funding, such as new cell lines or genetic databases, should "be made readily available for research purposes to qualified individuals within the scientific community."[215] These policies help insure that grant-funded research becomes available but consequently limit the availability of trade secrecy as a non-patent appropriation mechanism.

[209] *See, e.g.*, Ian Ayres & Lisa Larrimore Ouellette, *A Market Test for Bayh-Dole Patents*, 102 CORNELL L. REV. 271 (2017) (describing the inefficiency of the Bayh-Dole system and proposing a market mechanism for licensing of grant-funded inventions); Daniel J. Hemel & Lisa Larrimore Ouellette, *Bayh-Dole Beyond Borders*, 4 J.L. & BIOSCIENCES 282 (2017) (justifying the Bayh-Dole regime as useful to respond to challenges of global freeriding); Frischmann, *supra* note 31, at 399–413 (describing and critiquing the Bayh-Dole system of mixed grants and privately licensed patents); Stephen M. Maurer & Suzanne Scotchmer, *Procuring Knowledge*, *in* INTELLECTUAL PROPERTY AND ENTREPRENEURSHIP, 1, 26 (2004) (noting that if the Bayh-Dole Act solves any problem, it solves a problem with intellectual property law).

[210] *See* 35 U.S.C. § 203 (2018); 28 U.S.C. § 1498 (2018); *see also* Hannah Brennan et al., *A Prescription for Excessive Drug Pricing: Leveraging Government Patent Use for Health*, 18 YALE J.L. & TECH. 275 (2016) (describing the history of § 1498 and arguing that the federal government can use it today to buy generic versions of expensive drugs for far less than their list prices).

[211] Ayres & Ouellette, *supra* note 209, at 321; *see also* Ryan Whalen, *The Bayh-Dole Act & Public Rights in Federally Funded Inventions: Will the Agencies Ever Go Marching In?*, 109 NW. U. L. REV. 1083 (2015).

[212] *See* Brennan et al., *supra* note 210, at 280.

[213] *See supra* note 209.

[214] NIH GRANTS POLICY STATEMENT, *supra* note 70, at IIA-116.

[215] *Id.* at IIA-117; *see* Principles and Guidelines for Recipients of NIH Research Grants and Contracts on Obtaining and Disseminating Biomedical Research Resources, 64 Fed. Reg. 72090 (Dec. 23, 1999).

1. *Social Value Exceeding Market Value*

Grants can fund innovation whose social welfare value exceeds the market value of the product. Rachel Sachs notes,

> Where the general population's willingness and ability to pay for a particular drug track the social value it contributes, patents are thought to provide a relatively efficient way of incentivizing the development of socially valuable drugs. But each of these factors—willingness to pay and ability to pay—presents a well-known bias, through which innovation incentives will be directed away from certain types of treatments or diseases with high social salience.[216]

Willingness to pay creates a mismatch between social and private value. For some innovations, social benefits exceed individual benefits; for example, vaccines protect both the vaccinated individuals and others in society through the process of herd immunity.[217] Optimism bias may also decrease willingness to pay because people don't think they will get sick, and therefore underpay for preventive measures, decreasing incentives for scientists to develop those measures.[218] Finally, short-term bias may cause individuals to systematically undervalue expensive cures as opposed to ongoing treatments, which are cheaper per instance but costlier over time.[219] These distortions are not limited to the biomedical context—individuals may undervalue vehicle safety innovations that protect other drivers, upgrades that prevent house deterioration down the road, and technologies like solar roofs that pay for

[216] Rachel E. Sachs, *Prizing Insurance: Prescription Drug Insurance as Innovation Incentive*, 30 HARV. J.L. & TECH. 153, 168–69 (2016). Sachs suggests that insurance reimbursement may suggest another avenue to create incentives for this type of innovation. *Id.* at 178–93. Amy Kapczynski expands this argument generally in Amy Kapczynski, *The Cost of Price: Why and How to Get Beyond Intellectual Property Internalism*, 59 UCLA L. REV. 970 (2012).

[217] Sachs, *supra* note 216, at 169.

[218] *Id.* at 169–70 (citing Cass R. Sunstein, *Willingness to Pay vs. Welfare*, 1 HARV. L. & POL'Y REV. 303, 325 (2007)).

[219] *Id.* at 170. As Sachs notes, the story of Sovaldi, a drug which cures Hepatitis C, which itself primarily afflicts the poor, is somewhat miraculous. The typical story of drug market incentives suggests that firms should not be especially interested in a drug that treats a disease mostly afflicting those without substantial resources to pay, nor a drug that cures a chronic disease rather than treating it profitably for a long time. Sovaldi is both, and even its frequently-cited high sticker price represents a substantial savings over current treatment options. *See* Nicholas Bagley, *Does It Break the Law to Charge a Lot for a Cure?*, INCIDENTAL ECONOMIST (Jan. 28, 2016), http://theincidentaleconomist.com/wordpress/does-it-break-the-law-to-charge-a-lot-for-a-cure/ [https://perma.cc/32RS-RCXB] (quoting an email from Rachel Sachs to this effect). Outside biomedical innovation, climate change technology provides tremendous social benefits in the future, but current costs make appropriate market valuation of climate-change innovation challenging. *See generally* Ofer Tur-Sinai, *Patents and Climate Change: A Skeptic's View*, 48 ENV. L. REV. 211 (2018).

themselves in long-term energy savings.[220]

Ability to pay also limits market incentives and makes them inadequate for some socially valuable innovations. Consider Chagas, chikungunya, and other Neglected Tropical Diseases, which in the United States afflict mostly the poor and underinsured;[221] mental illness is similarly more prevalent among those populations.[222] Because those who can't pay for drugs can't create market demand, we should expect investment in treatments for those diseases to be substantially less than the social value of such innovation.[223]

These are not the only ways that market demand can create problematic incentives to pursue certain types of innovation. As Kevin Outterson has long argued, antibiotic resistance is a tremendous problem of global scale, caused in part by warped incentives for development of new antibiotics.[224] Antibiotic overuse limits the value of antibiotics for future users, but sellers of new antibiotics profit more from selling lots of the antibiotics before resistance sets in, rather than limiting their use.[225] Accordingly, new antibiotics aren't kept in reserve, and society loses the very large benefit of having a robust arsenal of last-resort antibiotics.[226] Unfortunately but perhaps unsurprisingly, the past several decades have seen little in the way of new antibiotics, and the looming threat of global antibiotic resistance is increasingly worrisome.[227]

Grants can step in to support research in these areas of unmet need. In 2012, for instance, the National Institute for Allergy and Infectious Diseases funded eight Tropical Medicine Research Centers through P50 Research Center grants.[228] The Centers are located in regions where the neglected tropical diseases are prevalent: Brazil, India, Ghana, and Peru.[229] These grants

[220] *See* Howard Kunreuther & Elke U. Weber, *Aiding Decision Making to Reduce the Impacts of Climate Change*, 37 J. CONSUMER POL'Y 397, 402–04 (2014).

[221] Sachs, *supra* note 216, at 154, 170–71.

[222] *Id.* at 170–71.

[223] *See, e.g.*, Stiglitz, *supra* note 39, at 1718 ("One of the problems of being poor is that you do not have any money and therefore cannot spend a lot of money on drugs, even though if you do not buy the drugs you may die.").

[224] *See generally* Kevin Outterson, *The Legal Ecology of Resistance: The Role of Antibiotic Resistance in Pharmaceutical Innovation*, 31 CARDOZO L. REV. 613 (2010).

[225] *Id.* at 627.

[226] *Id.*

[227] *See* Dalia Deak et al., *Progress in the Fight Against Multidrug-Resistant Bacteria? A Review of U.S. Food and Drug Administration-Approved Antibiotics, 2010–2015*, 165 ANNALS INTERNAL MED. 363, 369–71 (2016) (noting disappointing development of new antibiotics).

[228] *Tropical Medicine Research Centers – Program Overview*, NAT'L INST. ALLERGY & INFECTIOUS DISEASES, https://www.niaid.nih.gov/research/tmrc-program-overview [https://perma.cc/Q49Y-7W7L] (last visited Mar. 11, 2019).

[229] *Id.*

both support useful research in these areas of unmet need and "build capacity to enable [the Centers] to conduct future clinical trials, implement new treatment and prevention strategies, and develop novel vector control strategies."[230] In other words, the NIH aims to use grant funding to establish the capacity for future useful work even after initial funding has ended.[231] The Institute funds training to build research capacity, focusing on institutions in developing nations.[232]

Grants are unlikely to fully solve any of these problems of inadequate demand. But they provide useful innovation tools. Outterson and Aaron Kesselheim recognize that grants can play a role, within a complex system of tailored incentives, in supporting underlying research to reduce the cost of developing new antibiotics.[233] In a pleasant example of putting theory into practice, Outterson—in the years since he helped bring antibiotic resistance incentive problems to greater salience—has become the Executive Director of a $350-million grant-funded project aimed at increasing innovation in antibiotic development, including efforts that are too risky or paradigm-challenging for private development, as well as relatively mainstream efforts that suffer from the incentive problems described above.[234]

This type of grant-funding raises important questions: Who identifies underfunded innovations whose social value exceeds market value, and how? These questions may be especially challenging for applied research that does not obviously promote the same sorts of knowledge spillovers as basic research. Here, the first critique of grants—bureaucrats make funding decisions—has more bite.[235] But that may be precisely the point. This type of social welfare problem—social value that exceeds market price signals—is *exactly* the type of problem that market actors with private knowledge are ill-

[230] *Id.*

[231] *Id.*

[232] *See Funding Opportunity Announcement PAR-17-057: Global Infectious Disease Research Training Program (D43)*, NIH, https://grants.nih.gov/grants/guide/pa-files/PAR-17-057.html https://perma.cc/27K8-J49U] (last visited Mar. 11, 2019).

[233] Aaron S. Kesselheim & Kevin Outterson, *Fighting Antibiotic Resistance: Marrying New Financial Incentives to Meeting Public Health Goals*, 29 HEALTH AFF. 1689, 1694 (2010). Kesselheim and Outterson also respond to the risk allocation concern described *supra*, suggesting that "[f]or drugs that ultimately emerge from public investment programs, the government should receive an appropriate share of the enhanced reimbursement by payers." *Id.*

[234] *See* Kevin Outterson et al., *Accelerating Global Innovation to Address Antibacterial Resistance: Introducing CARB-X*, 15 NATURE REV. DRUG DISCOVERY 589 (2016).

[235] *See supra* Section II.B.1; *cf.* Abramowicz, *supra* note 50, at 1366–67 (arguing that orphan drug development should be subsidized only when they are inefficient and that government officials are likely unable to make that determination).

suited to fix.[236] The specialized knowledge of scientist peer reviewers might have more traction, but really, this is a problem about social welfare and identifying substantial unmet needs on a broader level. While the government (or philanthropic organizations) might do this inefficiently, it can make that social choice in a way that private firms won't.[237]

Even accepting that the government might be the right entity to make this sort of resource allocation call, how should it go about the task? Sachs argues that this sort of centralized decision-making is an opportunity for interagency collaboration to leverage different sources of knowledge and expertise.[238] With respect to under-addressed diseases, she notes that the Centers for Medicare and Medicaid Services (CMS) possess extensive information useful for NIH decisions on funding allocation, including on disease burdens; existing drugs; and, in combination, which diseases are currently underserved.[239] Unfortunately, such interagency collaboration is relatively underdeveloped,[240] but the collective expertise of CMS, the Centers for Disease Control, and other relevant agencies could help direct the funding allocation decisions of the NIH to address unmet biomedical needs with substantial potential social welfare gains.

2. *Appropriability Failures*

Grants can also pick up the incentive slack where markets value an innovation adequately, but firms cannot appropriate enough of its value to justify investment. The problem of appropriating the value of an information good is a fundamental justification for intellectual property. Ideally, intellectual property allows firms to appropriate social value of nonexclusive, nonrivalrous information goods by creating an exclusivity mechanism.[241] But intellectual

[236] In some cases, of course, no entity, whether private or public, will have a good answer as to the social value of a potential innovation. In such cases, whoever is making the decision must simply muddle through—as happens anyway. *See, e.g.*, Charles E. Lindblom, *The Science of "Muddling Through"*, 19 PUB. ADMIN. REV. 79 (1959) (explaining the difficulty in determining the social value of a policy).

[237] *Cf.* Pedraza-Fariña, *supra* note 10, at 439–41 (proposing that the grant system issue calls for scientists to propose important cross-disciplinary problems that need to be solved).

[238] Rachel E. Sachs, *Administering Health Innovation*, 39 CARDOZO L. REV. 1991, 1993–96 (2018).

[239] *Id.* at 2028.

[240] *See id.* at 2038–41; *see generally* Laura Pedraza-Fariña, *Constructing Interdisciplinary Collaboration: The Oncofertility Consortium as an Emerging Knowledge Commons*, *in* GOVERNING MEDICAL KNOWLEDGE COMMONS 259 (Kathy Strandburg, Brett Frischmann, & Michael Madison eds., 2017) (discussing failures in interagency collaboration in the context of the NIH Roadmap grants) [hereinafter Pedraza-Fariña, *Oncofertility*].

[241] *See generally* Mark A. Lemley, *Ex Ante Versus Ex Post Justifications for Intellectual Property*, 71 U. CHI. L. REV. 129 (2004) (discussing the different justifications that exist for having

property mechanisms don't always work. Where they fail, grants can step in, even if the innovation is relatively late in the development pipeline.[242]

To take one prominent example, medical diagnostics are a tough target for current patent law; grants could help. Diagnostics range from simple blood tests used in everyday care to the use of next-generation sequencing and complex multigene panels to pinpoint the cause of cancer. Often, the science underlying a diagnostic test is developed with grant funding. For instance, the Supreme Court's 2012 case about diagnostic methods patents, *Mayo v. Prometheus*, turned on a relationship between the proper dosing of a drug and the amount of a drug-related metabolite in the patient's blood.[243] That relationship was identified through grant-funded research, though the Court did not note that.[244] When the Court held in *Mayo* that the resulting diagnostic test was unpatentable as essentially stating a natural law (the underlying relationship) and telling doctors to "apply it,"[245] scholars (including me) noted that this description could cover many diagnostic tests, and worried that patents would no longer provide adequate incentives for firms to develop diagnostic tests and bring them into the market and into clinical use.[246] Some have suggested changing patent law to allay this concern.[247] But grants may do the job without needing to change patent law.[248]

Grants could support the process of bringing scientific relationships into use as diagnostic tests. For some diagnostics, not much needs to be done to go from relationship to test: once scientists identify genetic mutations associated with a disease (often using grant money), doctors can then identify

exclusive intellectual property).

[242] *See* Ouellette, *supra* note 22, at 1131–32, 1134–35, 1139.

[243] Mayo Collaborative Servs. v. Prometheus Labs, Inc., 566 U.S. 66, 73–75 (2012) (describing diagnostic technology in question).

[244] *See* Marla C. Dubinsky et al., *Pharmacogenomics and Metabolite Measurement for 6-Mercaptopurine Therapy in Inflammatory Bowel Disease*, 118 GASTROENTEROLOGY 705, 713 (2000) ("Supported by the Charles Bruneau Foundation . . . Fonds de la Recherche en Santé du Québec . . . and Fonds pour la Formation de Chercheurs et l'Aide à la Recherche . . . .").

[245] *Mayo*, 566 U.S. at 72–73.

[246] *See, e.g.*, Eisenberg, *supra* note 21; Sachs, *supra* note 21; W. Nicholson Price II, *Big Data, Patents, and the Future of Medicine*, 37 CARDOZO L. REV. 1401, 1425–26 (2016).

[247] *See, e.g.*, Jeffrey A. Lefstin, Peter S. Menell & David O. Taylor, *Final Report of the Berkeley Center for Law & Technology Section 101 Workshop: Addressing Patent Eligibility Challenges*, 33 BERKELEY TECH. L.J. 551 (2018) (outlining a workshop aimed at changing aspects of patent law).

[248] Changing patent law back to a pre-*Mayo* state would bring its own complications. *See, e.g.*, Price, *supra* note 246, at 1444–45 (briefly discussing these problems and citing more in-depth analyses). At a minimum, the Supreme Court seems uninterested in this possibility, having reaffirmed *Mayo* in *Alice*; change would require Congressional action.

the mutation after obtaining the patient's genetic sequence.[249] In those cases, additional grants may not even be needed. If more research needs to be done—exploring how well existing assays measure the relationship, whether the relationship accurately predicts status in various groups, and whether measurements can be used to improve clinical outcomes—grants can support this work without relying on patent incentives. And where doctors need new technology to apply newly discovered scientific relationships, patent law can still provide the market-calibrated incentives it does for other biomedical technologies—but focused on the technology, not the underlying relationship. For diagnostics, then, grants can support intermediate-cost technologies where some incentive is needed but other incentives are unavailable.

* * *

Grants are not unique in their ability to create incentives for innovation where social value exceeds appropriable market value. Prizes, in particular, can also provide incentives for such innovation, because they typically do not rely on exclusivity or matching market demand.[250] Indeed, prizes may work better in some circumstances where parallel effort between many research teams is demanded,[251] though they do not particularly help capital constrained firms.[252] R&D tax credits also create incentives for innovation where social value exceeds appropriable market value, though they do so by reducing innovation costs across the board rather than by targeting particular areas of likely social benefit. The point is not that grants are the only mechanism that can create incentives to solve this type of innovation problem, but that grants are a useful tool in this area, and that they use a different set of decision processes to create incentives. Grants *are* unique, however, in a different area.

[249] *See* Ass'n for Molecular Pathology v. U.S. Patent & Trademark Office, 653 F.3d 1329, 1338 (Fed. Cir. 2011) (describing role of DNA testing in diagnostics). In a parallel to *Mayo*, the Supreme Court held in *Myriad* that unaltered genomic DNA is unpatentable, making simple genetic tests of the "here's an important mutation; find it to diagnose a problem" variety similarly unpatentable. *See* Ass'n for Molecular Pathology v. Myriad Genetics, Inc., 569 U.S. 576, 580 (2013) (holding that "a naturally occurring DNA segment is a product of nature and not patent eligible merely because it has been isolated"). Today, doctors don't typically interpret genetic results on their own—genetic counselors act as intermediaries to interpret genetic testing results. A business model could rely on providing that intermediary service. *Cf.* Rachel E. Sachs, *Divided Infringement and the Doctor-Patient Relationship*, IP THEORY (forthcoming) (noting the difficulty of enforcing diagnostic methods patents in models with such intermediaries). But there is nothing to stop information about well-characterized mutations from becoming as routinely interpreted as, for instance, high cholesterol levels, once genetic sequencing becomes more common.

[250] *See* Adler, *supra* note 16, at 12–13.

[251] *Id.* at 13–14.

[252] *See* Hemel & Ouellette, *supra* note 2, at 336.

B. INNOVATION ENABLERS

Grants can support the people, institutions, processes, and infrastructure that enable innovation and shape its direction. Let me unpack that through a comparison with other innovation incentives. Patents focus on particular inventions: a patent protects the invention itself from appropriation by someone other than the patentee. Similarly, prizes address a particular product or outcome, like creating an accurate clock, finding a way to preserve food for a long period of time, creating a reusable vehicle for space flight, or the like.[253] Trade secrecy protects information, whether that be a way of manufacturing a challenging drug or a carefully assembled list of potential customers.[254] Each of these creates an incentive to develop the thing, the product, the output—and the rest of the innovation process is shaped around that incentive. Grants are different. They *can* focus on particular projects; indeed, many do. But grants can also fund individuals directly, allowing that individual to innovate in whatever way she sees best, whether that be toward a commercially viable product, basic knowledge production, or a set of several linked possibilities. Grants can aim squarely to build institutions, supporting centers or networks that can then pursue their own institutional research and innovation goals. They can shape innovation processes and build resources that enable fields to move forward. In this flexibility of focus, grants diverge sharply from patents, trade secrets, and prizes.[255] This Section describes four potential grant targets besides projects themselves: people, institutions, processes, and infrastructural datasets.

[253] *See, e.g.*, LONGITUDE PRIZE, https://longitudeprize.org/ [https://perma.cc/HL7C-YKN7] (last visited Mar. 11, 2019) (detailing the Longitude Prize, originally for ship navigation but currently for overcoming antibiotic resistance); Stephen Schaber, *Why Napoleon Offered a Prize for Inventing Canned Food*, NPR (Mar. 1, 2012) https://www.npr.org/sections/money/2012/03/01/147751097/why-napoleon-offered-a-prize-for-inventing-canned-food [https://perma.cc/LTE5-H9X4] (describing Napoleon's 1795 prize for improvement of food preservation methods); Tina Rosenberg, *Prizes with an Eye Toward the Future*, N.Y. TIMES (Feb. 29, 2012), https://opinionator.blogs.nytimes.com/2012/02/29/prizes-with-an-eye-toward-the-future/ [https://perma.cc/S8LZ-JY32] (noting the X Prize for private spaceflight); *see generally* Michael J. Burstein & Fiona E. Murray, *Innovation Prizes in Practice and Theory*, 29 HARV. J.L. & TECH. 401, 402–06 (2016) (describing the use of innovation prizes in general, including Longitude Prize and X Prize).

[254] *See* W. Nicholson Price II & Arti K. Rai, *Manufacturing Barriers to Biologics Competition and Innovation*, 101 IOWA L. REV. 1023, 1044–45 (2016) (explaining trade secrecy in relation to biologics manufacturing); *see also* Robert G. Bone, *The (Still) Shaky Foundations of Trade Secret Law*, 92 TEX. L. REV. 1803, 1805–06 (2014) (describing trade secret doctrine).

[255] Inasmuch as tax incentives create fungible incentives for any type of research undertaken by an entity which would otherwise owe income taxes, they function as an entity-targeted incentive rather than an outcome-focused incentive. *See* Hemel & Ouellette, *supra* note 2, at 321–26.

1. *People*

Two types of people might merit particular focus in terms of funding innovation: the exceptional and the young. Orthogonally, grants can support individuals either directly, without regard to project, or by weighing individual characteristics in addition to project merit.

a) The Exceptional and the Young

Why focus on exceptional individuals and the young? For the first, we might find it worthwhile to target truly exceptional individuals for grant support. That is, if we can identify the best scientists, we might judge them particularly good targets for grant funding because we think their projects are likely to be particularly influential.[256] We might also think them likely to produce, on average, more good ideas than other researchers; helping them pursue those projects rather than struggle for funding would increase social benefits.[257]

Separately, we might benefit from targeting younger scientists for grant support. Freeman and Van Reenen point to three reasons that younger scientists should be particular targets of grant funding: (1) in many fields, especially highly technical fields, researchers do their best work when they are relatively young, (2) providing funding early in a young scientist's career increases the odds that she will continue to pursue science, and (3) funding for scientists is among other things an investment in human capital.[258] All things being equal, a younger scientist has more time left in her career to use that capital (and to produce social benefits from that investment) than an older scientist.[259] Grant support is crucial to the careers of young scientists; as McGarity describes it, "[y]ounger scientists at prestigious institutions have no hope of becoming tenured if they do not have at least one NIH or NSF grant."[260]

[256] *See* Maurer & Scotchmer, *supra* note 209, at 17–18 (describing the need to identify the most creative individuals), 22–23 (arguing that researchers with the most fertile minds will self-select into the grant system).

[257] *Id.* at 23–24.

[258] For empirical evidence that human capital investments are more important to innovation than physical capital investments, see Fabian Waldinger, *Bombs, Brains, and Science: The Role of Human and Physical Capital for the Creation of Scientific Knowledge*, 98 REV. ECON. & STAT. 811, 811 (2016).

[259] Freeman & Van Reenen, *supra* note 91, at 22–23.

[260] McGarity, *supra* note 91, at 65 ("Denying a grant to a more established researcher can close his or her laboratory and effectively end his or her career as a productive researcher."); *see* Freeman & Van Reenan, *supra* note 91, at 19. Of course, grants are also important to later researchers.

Unfortunately, the current reality is that younger scientists have a hard time getting grant funding. The average age at which a PhD scientist gets her first R01 grant has been around 42 for several years; in 1980 it was 36.8.[261] Freeman and Van Reenen calculate that younger scientists have approximately tenfold worse chances of winning an R01 grant than scientists over 45.[262] This fact has worried scientists and policymakers, leading to policy changes including the mechanisms described in the next Section.[263]

b) Person-Focusing and Project-Weighting

A preference for a particular type of individual in grant funding can be implemented in at least two ways. First, grants can fund an individual separate from any project, to enable her to innovate, or to train her and therefore increase her human capital. Second, grant funding decisions can still focus on projects, but can heavily weight particular researcher characteristics.

Table 1: Examples of Grants Targeting People

	Person-focused		**Project preference**
	Training	*Enabling*	
Exceptional	n/a	HHMI, MacArthur	Implicit advantage
Young	F31, F32	n/a	ESI rules

Some individual-targeted grants focus entirely on enabling innovation by the individual. Training grants are common and aim to increase the expertise and human capital of the funded individual. The NIH offers several types of training grants, such as the F31 grant for supervised research training of doctoral candidates, the F32 grant for postdoctoral fellows "to broaden their scientific background and extend their potential for research," and the F33 Senior Fellow grant to help "experienced scientists to make major changes in the direction of research careers, or to acquire new research capabilities."[264] These grants are "training awards and not research awards."[265] They do not

[261] *See Average Age and Degree of NIH R01-Equivalent First-Time Awardees Fiscal Years 1980-2016*, NIH, https://grants.nih.gov/grants/new_investigators/Average_age_initial_R01.xls [https://perma.cc/63JU-9FNY] (last visited Mar. 11, 2019).

[262] Freeman & Van Reenen, *supra* note 91, at 21.

[263] *See, e.g.*, Ronald J. Daniels, *A Generation at Risk: Young Investigators and the Future of the Biomedical Workforce*, 113 PROC. NAT'L ACAD. SCI. 313 (2015) (describing the effects of declining research grants to young researchers to the biomedical industry).

[264] *Individual Fellowships*, NIH, https://researchtraining.nih.gov/programs/fellowships [https://perma.cc/GS28-HBB4] (last visited Mar. 11, 2019).

[265] NIH GRANTS POLICY STATEMENT, *supra* note 70, at IIB-37.

focus on the project, but rather the candidate's potential and need for training as well as how the proposed training, sponsor, and environment will address that need.[266] The K series grants similarly serve career development goals.[267] This group of grants focuses entirely on individuals and on enabling future innovation by building human capital.[268]

A different type of individual-enabling grant simply provides resources and an open mandate to an exceptional individual. The reasoning is that exceptional individuals, given freedom and resources, will tackle hard, risky problems and may produce exceptional results.[269] The NIH doesn't focus on this type of award, but other funders sometimes do. The Howard Hughes Medical Institute is perhaps the most substantial such funder and the MacArthur Foundation the closest follower of an individual-focused model. Howard Hughes, with the motto "People, Not Projects," identifies outstanding biomedical innovators, selects them as Howard Hughes Medical Investigators (currently there are around 300), and provides them with substantial funding—around $1 million per year—for renewable seven-year terms.[270] Howard Hughes aims to give "our scientists the time and freedom to pursue difficult, long-range questions,"[271] and at least some evidence suggests that this strategy works.[272] The MacArthur Foundation provides even purer grant funding to exceptional individuals, commonly known as Genius Grants.

[266] *Id.*

[267] *Id.* at IIB-80 (e.g., K01 grants for advanced research training and additional experience).

[268] *Id.*

[269] Patents can also highly reward the exceptional scientist, of course, but that depends on the research creating appropriable rewards; prizes depend on *post-hoc* recognition and typically do not provide funds to support research going forward.

[270] *See Fast Facts*, HOWARD HUGHES MED. INST., https://www.hhmi.org/press-room/fast-facts [https://perma.cc/A3WS-F2EZ] (last visited Mar. 11, 2019) (noting 292 current HHMI investigators); *HHMI Bets Big on 19 New Investigators*, HOWARD HUGHES MED. INST., https://www.hhmi.org/news/hhmi-bets-big-on-19-new-investigators [https://perma.cc/TAU3-UQ58] (last visited Mar. 11, 2019) (noting approximately $8 million in grants over a seven-year term for each Investigator). Technically, the researchers become HHMI employees, suggesting something more like a patronage model than classical grant funding. *See Our Scientists*, HOWARD HUGHES MED. INST., http://www.hhmi.org/scientists [https://perma.cc/DSS8-9J7C] (last visited Mar. 11, 2019). But they remain at their home institutions and retain their home appointments, and receive substantial funding to continue research in that context, making the appointment look very much like a person-focused grant. *Id.*

[271] *Biomedical Research Programs*, HOWARD HUGHES MED. INST., http://www.hhmi.org/programs/biomedical-research [https://perma.cc/5E8M-7Y9H] (last visited Mar. 11, 2019).

[272] *See* Azoulay, Zivin & Manso, *supra* note 202, at 528–29 (noting substantial differences in funding mechanisms and finding that Howard Hughes Medical Investigators produced more high-impact publications than NIH-funded scientists with similar accomplishments).

It provides "$625,000, no-strings-attached" five-year grants based on "[e]xceptional creativity" and the potential for substantial future work.[273] The Foundation "does not require or expect specific products or reports" from recipients.[274]

A different approach prioritizes projects by taking into account the characteristics of the individual researchers. The clearest example of this explicit prioritization comes in the NIH's special rules for grant applications by New and Early Stage Investigators—respectively, those who have not yet won a major research award and those within ten years of finishing their terminal degree.[275] For several years, the NIH has tried to reduce the age at which young scientists win their first major grants. The NIH clusters grant applications from New Investigators in peer review, so it can compare researchers with similar experience.[276] At least half of researchers receiving their first R01 or equivalent grant must be within ten years of finishing their terminal degree.[277] Finally, NIH Institutes make funding decisions aimed to achieve similar success rates for new grant applications by New Investigators and established investigators.[278] For instance, the National Cancer Institute's 2016 funding policy funded grants to the 10th percentile for established investigators but the 12th percentile for Early Stage Investigators—effectively putting a thumb on the scale for young researchers.[279] These policies generally reflect the goal of providing funding to younger scientists to invest in their futures—a goal that grants are uniquely suited to advance.[280]

[273] *About MacArthur Fellows Program*, MACARTHUR FOUND., https://www.macfound.org/programs/fellows/strategy/ [https://perma.cc/6L9N-L5BF] (last visited Mar. 11, 2019).

[274] *Id.*

[275] *Early Stage Investigator Policies*, NIH, https://grants.nih.gov/policy/new_investigators/index.htm [https://perma.cc/3HFY-DTL7] (last visited Mar. 11, 2019). For researchers who are medical doctors, Early Stage Investigators are those within ten years of finishing their medical residency. *See id.* To the best of my knowledge, no similar program exists for exceptional individuals—but exceptional researchers would be expected to submit exceptional grant applications in any case, and so should have an implicit advantage anyway.

[276] *See id.*

[277] *Id.*

[278] *See id.*

[279] NCI, *2016 Funding Strategy*, *supra* note 146. The NIH Director also has a set of grants to support extraordinary individuals, some of which, like the DP1 NIH Director's Pioneer Award, are specifically targeted at exceptional young researchers Grants. *See Types of Grant Programs*, NIH, https://grants.nih.gov/grants/funding/funding_program.htm [https://perma.cc/TZ3T-BDWV] (last visited Mar. 11, 2019) [hereinafter NIH, *Activity Codes*] (providing overview of NIH Activity Codes).

[280] Some prizes are explicitly targeted at the young, such as the Fields medal or the John Bates Clark medal, rewarded to outstanding mathematicians and economists, respectively,

1. *Institutions*

Grants can also target broader innovative entities, providing funding to institutions to enable future innovation. NIH grants typically fund research within a laboratory environment, whether in an academic institution, a hospital, or private industry. And indeed, NIH support for labs is critical; grants provide support for equipment, salaries, and research supplies, and are especially important in capital-constrained environments.[281] This ability of grants to purchase the equipment necessary for research has even been raised as a justification for the historical move from a prize-based to a grant-based innovation system.[282] More broadly, grants can enable the creation of new institutions, support the efforts of existing institutions, or allow existing institutions to increase their capacity. Similar to the focus on exceptional individuals described above, the NIH can identify institutions that are likely to be especially productive and help them increase their capabilities.

The NIH provides many grants specifically targeted at increasing institutional capabilities. The G11 grant helps institutions improve their research infrastructure by providing funds for them to establish an office of sponsored research to work with grant funders.[283] M01 grants support "General Clinical Research Center[s] where scientists conduct studies on a wide range of human diseases using the full spectrum of the biomedical science," and can fund renovation, staff salaries, equipment, and supplies.[284] P01 grants support research programs, P30 grants support administrative cores for centers, P51 grants support primate research colonies, and P60 grants support comprehensive centers—the list goes on.[285] Suffice it to say, the NIH can and does target institutions, centers, and programs of different sizes and foci, all to further the goal of enabling innovation by those best suited to innovate. As with focusing on individuals, this institution-supporting role is essentially unique to grants.

under the age of forty. *Fields Medal*, INT'L MATHEMATICAL UNION, https://www.mathunion.org/imu-awards/fields-medal [https://perma.cc/4VSW-EWA4] (last visited Mar. 11, 2019); *John Bates Clark Medal*, AM. ECON. ASS'N, https://www.aeaweb.org/about-aea/honors-awards/bates-clark [https://perma.cc/5LG3-DQU4] (last visited July 2, 2017). However, such prizes generally do not provide substantial funds for either training or research going forward.

[281] Hemel & Ouellette, *supra* note 2, at 334–38.

[282] *See, e.g.*, Hanson, *supra* note 52, at 7–8.

[283] *See* NIH, *Activity Codes*, *supra* note 279 (G11 grant description available from dropdown list).

[284] *Id.* (M11 grant description available from dropdown list).

[285] *Id.*

2. *Processes*

Grants can influence the processes through which innovation takes place; in particular, they can create incentives for collaboration and interdisciplinary work. Again, this focus differs from other incentives; patents, prizes, and tax R&D incentives tend not to take account of innovation environment. Collaboration may impact the value of these rewards—joint inventorship changes the control mechanisms for patents, and of course joint creation splits the reward of any of these mechanisms—but other policy levers do not specifically encourage collaboration.[286] Grants, by contrast, can and do.

Grants can generally target collaborative work where researchers from different labs or institutions work together on a funded project. Encouragement can be explicit, such as requirements that recipients participate in collaborative research networks.[287] Elias Zerhouni, the Director of the NIH, launched the 2002 Roadmap for Medical Research Initiative specifically to encourage and fund collaborative team science.[288] Grants may also implicitly encourage collaboration by preferentially funding projects that require collaborative work.[289]

An important subset of process-focused grants promotes interdisciplinary work. Boundary-crossing work can push forward the frontiers of science and innovation.[290] However, interdisciplinary work is hard; it is challenging to master multiple disciplines or to reach across disciplinary lines, and interdisciplinary researchers may encounter resistance from peers and scientific institutions.[291] Such work is also "high-risk, high-reward," suggesting

[286] *See* Gregory N. Mandel, *To Promote the Creative Process: Intellectual Property Law and the Psychology of Creativity*, 86 NOTRE DAME L. REV. 1999, 2001 (2011) ("Problematically, the laws of joint authorship and joint inventorship in intellectual property actually dissuade certain collaboration.").

[287] Interview with Anonymous Senior Scientist (June 7, 2017) (on file with author) (describing grant requirement that recipients participate in a research network and noting that it led to productive collaborative work).

[288] *See generally* Elias A. Zerhouni, *The NIH Roadmap*, 302 SCIENCE 63 (2003).

[289] *See* Robin Barr, *R01 Teams and Grantee Age Trends in Grant Funding*, NIH NAT'L INST. ON AGING (April 22, 2015), https://www.nia.nih.gov/research/blog/2015/04/r01-teams-and-grantee-age-trends-grant-funding [https://perma.cc/F7JB-6EWL] (noting that the modal top-scoring R01 grant in 2005 had one principal investigator; in 2015 it had four).

[290] *See* JULIE THOMPSON KLEIN, INTERDISCIPLINARITY: HISTORY, THEORY, AND PRACTICE 12 (1990); Pedraza-Fariña, *supra* note 10, at 439–41; *see also* Michal Shur-Ofry, *Connect the Dots: Patents and Interdisciplinarity*, 51 MICH. J.L. REFORM 55, 62–65 (2017).

[291] *See* Pedraza-Fariña, *supra* note 10, at 423–24 (discussing social barriers to interdisciplinary innovation). There is a rich literature outside law on interdisciplinarity. *See, e.g.*, Susan Leigh Star & James R. Griesemer, *Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39*, 19 SOCIAL

that innovation incentives are likely to be useful in promoting investment. Unfortunately, patents aren't especially good at promoting interdisciplinary work; Michal Shur-Ofry writes that patent law generally regards interdisciplinary combinations "not as a potential source of groundbreaking innovation, but at most, as an excusable flaw."[292]

Grants, on the other hand, can directly target and facilitate interdisciplinary work.[293] Laura Pedraza-Fariña examines an NIH grant program that aimed squarely at interdisciplinary work.[294] She focuses on one part of Zerhouni's Roadmap, the Interdisciplinary Research Consortia grants, which funded nine interdisciplinary consortia between 2005 and 2012.[295] Pedraza-Fariña recounts the formation of the Oncofertility Consortium, a network of researchers focused on solving the problem of oncofertility—that is, how can we ensure that cancer patients can still have children after their treatment?[296] Oncofertility is a knotty scientific problem, and a tough interdisciplinary one: oncologists, reproductive endocrinologists, and basic research scientists have substantially different approaches and areas of expertise.[297] Pedraza-Fariña describes how the grant program, which specifically called for interdisciplinary applications, served as a "catalyst to collaboration—providing short-term, seed funding to enable cross-disciplinary collaboration."[298] It did so by combining several different grant types, including some types described above: a U54

STUD. SCI 387 (1989) (coining the term "boundary object"); TRADING ZONES AND INTERACTIONAL EXPERTISE: CREATING NEW KINDS OF COLLABORATION (Michael E. Gorman ed., 2010) (discussing framework for fostering interdisciplinary collaborations).

[292] Shur-Ofry, *supra* note 290, at 72; *see* Pedraza-Fariña, *supra* note 10, at 436–38 (arguing that patent doctrine is actively hostile to interdisciplinary innovation and suggesting modifications); Mandel, *supra* note 286; Jacob S. Sherkow, *Negativing Invention*, 2011 BYU L. REV. 1091, 1094–95 (2011) (noting that interdisciplinary combinations are less susceptible to "analogous arts," and have the effect of "negativing" inventions).

[293] *See* Pedraza-Fariña, *supra* note 10, at 442 ("[G]overnment grants or prizes can be structured to incentivize the identification of problems whose solution requires the combined expertise from multiple disciplines and subdisciplines."). Note that collaboration and interdisciplinarity are not targets only of NIH grants, nor indeed only of federal grants; they can be targeted by any grant funder. *See, e.g.*, *MCubed*, UNIV. MICH., *supra* note 83 (noting that funding will be provided only to teams of at least three faculty researchers from at least two different campus units).

[294] *See generally* Pedraza-Fariña, *Oncofertility*, *supra* note 240.

[295] *Id.* at 260 (citing *Interdisciplinary Program Snapshot*, NIH, https://commonfund.nih.gov/Interdisciplinary [https://perma.cc/8U8V-JJ28]). Because obtaining cross-disciplinary grants from individual disease-focused NIH Institutes is hard, the broader Interdisciplinary Research program was funded by the Common Fund, a central pool of money used for larger strategic NIH initiatives. *Id.*

[296] *Id.* at 260.

[297] *Id.* at 260–61.

[298] *Id.* at 261.

Cooperative Agreement for a specialized center to support a centralized administrative core to organize and coordinate the team,[299] four R01 Research Project grants to support basic research into female follicles,[300] two P30 Center Core grants to fund a core for maintaining and distributing patient samples and other materials and to fund the National Physician's Cooperative (the network of participants), an R25 Education Project grant to fund an "educational module," and three different grants (a T90 Interdisciplinary Research Training Award, an R90 Interdisciplinary Regular Research Training Award, and a K01 Research Scientist Development Award - Research & Training) to fund training for oncofertility specialists.[301] The Interdisciplinary Research Consortia program leveraged several different grant regimes with the goal of not only supporting interdisciplinary collaboration, but also of catalyzing something that would last long-term. In short, the grant program tried to use a jolt of focused funding to create something novel and sustainable.

And it worked. As Pedraza-Fariña documents, the Oncofertility Consortium developed specifically in response to the Interdisciplinary Research Consortia program's call for applications. Although the scientists involved knew each other, "none of them . . . had embarked on a collaboration of this magnitude, nor held a focused discussion on how to address fertility preservation questions in a concerted manner prior to applying for the oncofertility consortium grant."[302] Although the Interdisciplinary Research Consortia program ended in 2012, the Oncofertility Consortium continues today.[303] In addition, the Consortium has built infrastructure that can be used going forward and has spawned other ongoing collaborations.[304]

The Oncofertility Consortium was not the only interdisciplinary consortium funded by the NIH's program. The program also funded consortia focused on the molecular mechanisms of stress; the science of aging with a focus on cancer, organ design and engineering; and obesity and metabolic disorders, among others.[305] At least some are still active today.[306] And as

[299] *Id.* at 280.

[300] *Id.* at 262 n.20.

[301] *Id.* at 280.

[302] *Id.* at 275.

[303] THE ONCOFERTILITY CONSORTIUM, http://oncofertility.northwestern.edu/ [https://perma.cc/B493-SUKY] (last visited Mar. 11, 2019).

[304] Pedraza-Fariña, *Oncofertility*, *supra* note 240, at 283.

[305] *See Interdisciplinary Research Consortia*, NIH, https://commonfund.nih.gov/Interdisciplinary/consortia [https://perma.cc/FM58-MWH3] (last visited Mar. 11, 2019).

[306] In comparison with the still-vital Oncofertility Consortium, see, for example, *Taskforce for Obesity Research at UT Southwestern (TORS)*, U.T. SOUTHWESTERN, https://www.utsouthwestern.edu/education/medical-school/departments/center-human-

Pedraza-Fariña points out, overcoming initial hurdles to collaboration may be much of the battle; even if particular consortia end, the possibility of interdisciplinary collaboration remains easier after the initial structural work has been done—work that grants can specifically target and support.[307]

3. *Infrastructure*

Finally, grants can specifically target infrastructural goods to create broad support for future innovation. Brett Frischmann characterizes infrastructural goods by three key traits: (1) they "may be consumed nonrivalrously for some appreciable range of demand"; (2) they are valuable largely because they are inputs into downstream productive activities; and (3) such activities may produce a wide range of goods, including public goods, social goods, and private goods.[308] Infrastructural goods are socially valuable because they enable a broad range of activities and have many spillovers; they are public goods and enable others to generate public goods.[309] But that's why the incentives to invest in infrastructural goods tend to be too low. On the supply side, it is hard for infrastructure investors to appropriate the full social benefits of their investment: infrastructure has spillover benefits that are hard to capture.[310] And on the demand side, even if infrastructure investors could appropriate all the private demand for the infrastructural good, users are unlikely to be willing to *pay* the full social value for access to the infrastructure, because they may be creating public goods whose benefits *they* cannot appropriate.[311] All of which is to say: infrastructural goods have substantial social benefits, but it is rare for private entities to have the right incentives to either create the infrastructure in the first place or allow broad enough, cheap enough use that downstream users create the largest social value.[312]

Enter grants. The government can get involved to help overcome the challenges with private incentives for infrastructure.[313] Sometimes that is direct construction; for example, the federal government built and runs the interstate highway system.[314] Sometimes not; grants can provide a powerful way to

nutrition/obesity-alliance.html [https://perma.cc/LF23-VR2X] (last visited Mar. 11, 2019) (showing no publications after 2012 and no conference meetings after 2014).

[307] Pedraza-Fariña, *Oncofertility*, *supra* note 240, at 283–84.

[308] BRETT M. FRISCHMANN, INFRASTRUCTURE: THE SOCIAL VALUE OF SHARED RESOURCES 61–62 (2013).

[309] *Id.* at 68–69.

[310] *Id.* at 14–15.

[311] *Id.* at 71–72.

[312] *See id.* at 98.

[313] *Id.* at 14–15.

[314] *Id.* at 189–90.

leverage non-governmental expertise in large, infrastructural projects designed to create resources that will broadly enable future scientific endeavors.[315] These projects tend to be motivated by the centralized belief—held by both administrators and scientists—that the infrastructure project will create substantial social value. Prominent NIH programs have thus used grants to drive large-scale scientific infrastructure projects and to make their fruits broadly available.[316]

The Human Genome Project, which started at the end of the 20th century, is a key example.[317] The Project was a massive undertaking that aimed to sequence the entire human genome.[318] The explicit goal of the project was to create infrastructure for future research, "to provide researchers with powerful tools to understand the genetic factors in human disease, paving the way for new strategies for their diagnosis, treatment and prevention."[319] Liscow and Karpilow highlight the potential for government spending to shift the course of future innovation: where legacy technologies (in their example, high-pollution fossil fuel technology) benefit from a large existing stock of knowledge, concentrated government efforts to support knowledge generation in a new technology can shift future innovation in a socially desirable direction.[320]

The Human Genome Project followed this pattern, creating benefits beside the genome map itself. The production of a human genome sequence enabled a large set of downstream uses, including developments in pharmacogenomics and genetic testing.[321] It helped shift innovation away from

[315] *See generally* Jorge L. Contreras, *Leviathan in the Commons: Biomedical Data and the State*, *in* GOVERNING MEDICAL KNOWLEDGE COMMONS 19 (Kathy J. Strandburg, Brett M. Frischmann & Michael J. Madison eds., 2017) (describing the ways government actors shape biomedical data resources beyond merely supporting their creation).

[316] *See* Jorge L. Contreras, *Bermuda's Legacy: Policy, Patents, and the Design of the Genome Commons*, 12 MINN. J.L. SCI. & TECH. 61 (2011) (describing the evolution of data release policies for genomic data starting with the 1996 Bermuda Principles); Jorge L. Contreras, *Constructing the Genome Commons*, *in* GOVERNING KNOWLEDGE COMMONS 99, 102 (Brett M. Frischmann, Michael J. Madison & Kathy J. Strandburg eds., 2014) (describing genomic data as a commons with a "unique polycentric governance institution").

[317] *See* HILGARTNER, *supra* note 174 (describing the history of the Genome Project, focusing on the creation and change of knowledge-control regimes).

[318] Of course, there is no one human genome; almost everyone's is different. The Project aimed to generate a generalized consensus sequence upon which variations could be mapped.

[319] NIH, FACT SHEET: HUMAN GENOME PROJECT 1 (Oct. 2010), https://report.nih.gov/NIHfactsheets/Pdfs/HumanGenomeProject(NHGRI).pdf [https://perma.cc/NY3D-2XEQ] [hereinafter NIH, FACT SHEET].

[320] Liscow & Karpilow, *supra* note 51, at 392–93.

[321] *See* NIH, FACT SHEET, *supra* note 319 (noting thousands of disease genes discovered, thousands of new genetic tests, hundreds of biotechnology products in clinical trials, and

the use of inexact or problematic proxies, like using race as a proxy for unmeasured genetic traits, and toward more direct genetic diagnostics.[322] The Human Genome Project also created a guaranteed demand for technological advances that otherwise might be too risky, including novel genetic sequencing technology.[323] The project explicitly sought to develop technology and information infrastructure, eventually leading to lower costs despite the initial outlay.[324]

A private effort to sequence the human genome, Celera Genomics, illustrates the role of the government in such infrastructure projects. Celera Genomics entered the fray several years after the Human Genome Project began, aiming to complete its sequence much faster than the publicly funded effort.[325] But Celera Genomics' own effort—while impressive, fast, and generating and leveraging its own technological advances—itself relied substantially on publicly funded sequence data infrastructure resources.[326] According to Steven Hilgartner's history of the Human Genome Project, approximately 60% of the completed sequence shared in Celera's Science paper was in fact downloaded from the Human Genome Project's publicly available dataset.[327] The differences between the two projects also illuminate the benefits of publicly funded, relatively open management of infrastructural resources.[328] The publicly funded effort helped develop the technology that supported the private effort—which then developed its own tremendously useful technology and created an important comparator sequence.[329] But even once both sequences existed, Celera's management of its own sequence as a private resource with paid access limited the sequence's uses to those with the resources to pay, and, likely, to a subset of uses with more potential for immediate commercial gain rather than basic research or other projects with

ongoing enabled scientific research).

[322] *See* W. Nicholson Price II, Note, *Patenting Race: The Problems of Ethnic Genetic Testing Patents*, 8 COLUM. SCI. & TECH. L. REV. 119, 134–37 (2007).

[323] *Cf.* Glennerster, Kremer & Williams, *supra* note 57, *passim* (describing advance purchase commitments as a mechanism to create incentives for firms to develop vaccines that otherwise might be too risky to draw enough investment).

[324] *See* HILGARTNER, *supra* note 174, at 50.

[325] *Id.* at 206–10.

[326] *Id.* at 221.

[327] *Id.*

[328] *See supra* notes 308–313 and accompanying text.

[329] *See* Int'l Human Genome Sequencing Consortium, *Initial Sequencing and Analysis of the Human Genome*, 409 NATURE 860 (2001) (announcing the Human Genome Project's completed sequence); *see also* J. Craig Venter et al., *The Sequence of the Human Genome*, 291 SCIENCE 1304 (2001) (announcing Celera's completed sequence).

greater spillovers.[330]

Today, grants can help develop precision medicine and complex algorithms based on medical big data. The Precision Medicine Initiative aims to advance our knowledge of precision medicine, providing "the right treatments to the right patients at the right time."[331] It does this by supporting basic scientific research along these lines and also by partnering with many institutions to gather extensive genetic and health information, as well as biospecimens, on over one million volunteers—the "All of Us" cohort—as an infrastructural resource for future innovation.[332] As Sachs notes, such large infrastructural initiatives can also focus other stakeholder efforts; the Precision Medicine Initiative has stimulated non-governmental investors to commit over $200 million.[333]

A step further in the future, complex medical algorithms have the potential for tremendous benefits to the health care system, including improving patient care, optimizing resource allocation, suggesting new possibilities for treatment, and identifying problems or unknown benefits of existing drugs.[334] But current innovation incentives are problematic. Patents are often unavailable, and relying on secrecy for databases or algorithms creates an array of problems.[335] In addition, market signals of demand may substantially underrepresent social value, particularly for the collection and use of data for underserved populations, including poor and minority populations.[336] NIH grants could support the development of infrastructure, focusing on assembling and curating data, especially for underserved populations, and making it broadly

[330] HILGARTNER, *supra* note 174, at 212–13.

[331] *See Precision Medicine*, U.S. FOOD & DRUG ADMIN., https://www.fda.gov/medicaldevices/productsandmedicalprocedures/invitrodiagnostics/precisionmedicine-medicaldevices/default.htm [https://perma.cc/9ZMZ-GK2E] (last visited Mar. 11, 2019).

[332] *Scientific Opportunities*, NIH, https://allofus.nih.gov/about/scientific-opportunities [https://perma.cc/H7EX-6V6A] (last visited Mar. 11, 2019) ("The program will set the foundation for new ways of engaging research participants, sharing health data and information, and employing technology advances to mine the information for comprehensive results."); *see Awardees*, NIH, https://allofus.nih.gov/funding/awardees [https://perma.cc/PA8E-TZNW] (last visited Mar. 11, 2019) (noting award of a U24 Cooperative Agreement to the Mayo Clinic to host a specimen biobank).

[333] Sachs, *supra* note 238, at 2002 (citing Press Release, White House, FACT SHEET: Obama Administration Announces Key Actions to Accelerate Precision Medicine Initiative (Feb. 25, 2016)).

[334] W. Nicholson Price II, *Black-Box Medicine*, 28 HARV. J.L. & TECH. 419, 434–37 (2015) (discussing "black-box medicine" and the use of opaque computational models to make decisions related to health care).

[335] Price, *supra* note 246, at 1419–36.

[336] *See supra* Section IV.A.1.

available to the research community.[337] The "All of Us" cohort provides a start, but the NIH could go even further, broadening the reach of potential data, populations included, and research analyses supported.

V. CONCLUSION

Grants are a useful tool in the innovation toolbox. Although there is some truth to common critiques—bureaucrats are involved in decisions, individual grants are indeed ex ante funding with relatively low accountability, and the government doesn't profit directly from grant-funded research—the reality is much more complex, and the critiques mask this complexity. Funding decisions are largely made by scientists based on scientific merit, the repeat-player nature of grants creates accountability, and the government and society reap substantial indirect benefits from grants whether they succeed or fail. Moreover, the grant application process and the peer review process bring considerable information and expertise to bear on government choices about what projects to fund.

While I am enthusiastic about what the grant system has to offer, I do not mean to suggest, a naïve Pollyanna, that the system is wrinkle-free. The three critiques have some truth to them and the system has other problems. The system of repeat players can privilege experience and erect barriers to entry for new innovators, especially innovators who do not tread the typical path.[338] Seeking grants can consume inordinate amount of a researcher's time and energy;[339] postdoctoral fellows can be trapped in fellowships or chased from science by the unavailability of grants.[340] And the hunt for scarce money can warp research priorities despite the best efforts of funders and peer reviewers. Grants are not perfect.

Nevertheless, the overall system, the aggregation of scientific knowledge

[337] *See, e.g.*, W. Nicholson Price II, *Risk and Resilience in Health Data Infrastructure*, 16 COLO. TECH. L.J. 65, 77–83 (2017) (discussing the benefits of investment in health data infrastructure).

[338] *See supra* notes 158–165 and accompanying text.

[339] *See, e.g.*, *Dr. No Money: The Broken Science Funding System*, SCI. AM. (May 1, 2011), https://www.scientificamerican.com/article/dr-no-money/ [https://perma.cc/TRB9-G5ML] (arguing that scientists spend too much time raising funds instead of doing experiments); Matt Welsh, *The Secret Lives of Professors*, MATT-WELSH.BLOGSPOT (May 24, 2010), http://matt-welsh.blogspot.com/2010/05/secret-lives-of-professors.html [https://perma.cc/NCQ6-FLF8] (discussing the marketing and fundraising aspect of science).

[340] *See, e.g.*, Muhammed Z. Ahmed, *The Postdoc Crisis*, SCIENTIST (Jan. 4, 2016), https://www.the-scientist.com/opinion/opinion-the-postdoc-crisis-34259 [https://perma.cc/6WQW-KKSN] (arguing that postdoctoral fellows have few prospects in academia because of funding issues).

and priorities—with input from the government as to social benefit—is not inferior to determinations that arise from private market aggregation of private knowledge; it's just different. The grant system has its own flaws and foibles, but also, importantly, presents an alternative decision-making process that avoids the flaws and foibles of the market-dominated systems of other innovation levers. If our only goal is the cheapest development of drugs for the wealthy, then we can probably rely only on market mechanisms to allocate innovation and do just fine. But if we care more broadly about the formation of new scientific fields before the promise of obvious commercial profits, the development of drugs for the poor, the creation of difficult-to-exclude knowledge, the nourishment of mobile young scientists, the creation of interdisciplinary networks, or the pursuit of other goals that the market and private knowledge can neither appropriately value nor staff, then grants provide an attractive set of policy options. Grants are not the only way to pursue these goals, but they use a different way of gathering information and allocating resources that make such pursuits more straightforward.

A complete understanding of the role of grants in the innovation ecosystem demands more study, both theoretical and empirical. In addition to comparisons of grants with other innovation levers that incorporate a more nuanced view of grants, future studies could examine more closely how different levers function together.[341] Innovation levers don't work in a vacuum; trade secrets exist before patents, researchers can patent results of both grant-funded research and private research subsidized through the tax system, prizes kick in at the end, and grants can stretch across multiple innovative efforts. We should understand how these levers work in concert—or how they compete against and distort one another.[342] Such scholarship could include large-scale quantitative analyses of many actors across the economy, small-scale examinations of specific innovation contexts,[343] or theoretical conceptions of how different levers can and should interact.[344] The political

[341] Brett Frischmann, Michael Madison, and Kathy Strandburg's work on studying innovation commons involves this sort of thick, cross-lever innovation exploration, though focused on the role of information commons. *See generally* GOVERNING MEDICAL KNOWLEDGE COMMONS (Katherine J. Strandburg, Brett M. Frischmann & Michael J. Madison, eds., 2017).

[342] *See, e.g.*, Price & Rai, *supra* note 254 (describing innovation-stifling effects from the intersection of patents, trade secrecy, and regulatory product definitions).

[343] *See, e.g.*, Sarnoff, *Likely Mismatch*, *supra* note 31, at 374–80 (noting the context specificity of innovation incentives); Gallini & Scotchmer, *supra* note 38 (same); Hemel & Ouellette, *supra* note 2, at 378–80 (discussing the mix of innovation levers deployed in the context of orphan drugs).

[344] *See, e.g.*, Daniel J. Hemel & Lisa L. Ouellette, Innovation Policy Pluralism (2017) (unpublished manuscript) (on file with author) (theorizing and describing examples of the

economy of grants—routinely receiving bipartisan support from Congress, but nonetheless vulnerable to political vicissitudes and potentially changing funding[345]—further shapes their place in the innovation policy toolbox and deserves closer examination in this literature. Finally, studies of grants as part of the innovation policy toolbox should consider the nitty-gritty details of how grants work best on the ground, incorporating empirical studies from the economics of innovation into the design of research policy.[346] Improving grant functioning could even involve its own experimentation, changing funding mechanisms for just a subset of innovators and evaluating the results.[347] Grants are a key part of the innovation ecosystem, but they are often not treated that way by the literature on innovation law and policy. It is time for that to change.

mixture of intellectual property and non-IP mechanisms in innovation policy).

[345] *See* Deepak Hegde & David C. Mowery, *Politics and Funding in the U.S. Public Biomedical R&D System*, 322 SCIENCE 1797 (2008) (noting some evidence of the politicization of the grants process); Pear, *supra* note 5 (reporting that Congress rejected President Trump's proposal to cut N.I.H funding and instead increased funding).

[346] *See, e.g.*, Freeman & Van Reenan, *supra* note 91 (examining the impact of the 1998–2003 doubling of the NIH budget on the biomedical sciences); Michael Levitt & Jonathan M. Levitt, *Future of Fundamental Discovery in US Biomedical Research*, 114 PROC. NAT'L ACAD. SCI. 6498 (2017) (finding bias against awarding grants to younger applicants, in favor of older principal investigators).

[347] *See* Pierre Azoulay, Joshua S. Graff Zivin & Gustavo Manso, *National Institutes of Health Peer Review: Challenges and Avenues for Reform*, *in* 13 INNOVATION POLICY & THE ECONOMY 1, 13–16 (Josh Lerner & Scott Stern eds., 2013) (examining peer-review practices in light of NIH's bias for funding older scientists and the innovativeness of that funded research).

ALGORITHMS AS ILLEGAL AGREEMENTS

Michal S. Gal[†]

ABSTRACT

Algorithms offer a legal way to overcome some of the obstacles to profit-boosting coordination, and create a jointly profitable status quo in the market. While current research has largely focused on the concerns raised by algorithmic-facilitated coordination, this Article takes the next step, asking to what extent current laws can be fitted to effectively deal with this phenomenon. To meet this challenge, this Article advances in three stages. The first Part analyzes the effects of algorithms on the ability of competitors to coordinate their conduct. While this issue has been addressed by other researchers, this Article seeks to contribute to the analysis by systematically charting the technological abilities of algorithms that may affect coordination in the digital ecosystem in which they operate. Special emphasis is placed on the fact that the algorithms is a "recipe for action," which can be directly or indirectly observed by competitors. The second Part explores the promises as well as the limits of market solutions. In particular, it considers the use of algorithms by consumers and off-the-grid transactions to counteract some of the effects of algorithmic-facilitated coordination by suppliers. The shortcomings of such market solutions lead to the third Part, which focuses on the ability of existing legal tools to deal effectively with algorithmic-facilitated coordination, while not harming the efficiencies they bring about. The analysis explores three interconnected questions that stand at the basis of designing a welfare-enhancing policy: What exactly do we wish to prohibit, and can we spell this out clearly for market participants? What types of conduct are captured under the existing antitrust laws? And is there justification for widening the regulatory net beyond its current prohibitions in light of the changing nature of the

DOI: https://doi.org/10.15779/Z38VM42X86

[†] Professor and Director of the Forum for Law and Markets, University of Haifa Faculty of Law; President, International Association of Competition Law Scholars (ASCOLA). Many thanks to Ariel Ezrachi, Joe Harrington, Steve Tadelis, and Bill Page, and the editors of the law review, who provided excellent comments on previous drafts, and to Bill Kovacic, Nicolas Petit, and Moshe Tenenholtz for thought-provoking discussions. I also benefitted from comments at the OECD Panel on Algorithms and Collusion, the Oxford Conference on Digital Markets, the UK Competition and Markets Authority Distinguished Speaker Series, and the New Zealand Commerce Commission's bi-Annual Conference, and participants in seminars in Harvard, NYU and Columbia Universities, for most helpful discussions and/or comments on previous drafts. Ido Shapira and Ilana Atron provided excellent research assistance. This research was supported by the Center for Cyber Law and Policy. Any mistakes or omissions remain the author's. The paper was partly written while I was a Visiting Professor at Columbia University. I thank them for their warm hospitability and invigorating environment. I would like to give special thanks to the editors of BTLJ for their hard work and support in bringing this Article to publication.

marketplace? In particular, the Article explores the application of the concepts of plus factors and facilitating practices to algorithms. The analysis refutes the claim that current laws are sufficient to deal with algorithmic-facilitated coordination.

TABLE OF CONTENTS

"We will not tolerate anticompetitive conduct, whether it occurs in a smoke-filled room or over the Internet using complex pricing algorithms . . . . Consumers have the right to a free and fair marketplace online, as well as in brick and mortar businesses."[1]

I. INTRODUCTION

Despite the increased transparency, connectivity, and search abilities that characterize the digital marketplace, the digital revolution has not always yielded the bargain prices that many consumers expected. Why not? Some researchers suggest that one factor may be coordination between the algorithms that are used by suppliers to determine trade terms.[2] Coordination-facilitating algorithms are already available off the shelf, and such coordination is only likely to become more commonplace in the near future. This is not

[1] Press Release, U.S. Dep't of Just. Antitrust Div., Former E-Commerce Executive Charged with Price Fixing in the Antitrust Division's First Online Marketplace Prosecution (Apr. 6, 2015) (quoting Assistant Attorney General Bill Baer).

[2] *See infra* Section II.B.

surprising. If algorithms offer a legal way to overcome obstacles to profit-boosting coordination, and to create a jointly profitable status quo in the market, it is no surprise that suppliers use them. In light of these developments, seeking solutions to algorithm-driven coordinated high prices—both regulatory and market-driven—is timely and essential. While current research has largely focused on the concerns raised by algorithmic-facilitated coordination, this Article takes the next step, asking to what extent current laws can be fitted to effectively deal with this phenomenon.

The use of algorithms in digital markets creates many benefits. Algorithms allow consumers to efficiently compare products and offers online, enabling them to enjoy lower-priced goods or find products that better fit their preferences.[3] Suppliers can quickly and efficiently analyze large amounts of data, allowing them to better respond to consumer demand, better allocate production and marketing resources, and save on human capital.[4] To achieve these results, algorithms perform a myriad of tasks, including collecting, sorting, organizing and analyzing data, making decisions based on that data, and even executing such decisions.

Some of these advantages are currently threatened by algorithmic-facilitated coordination among competitors.[5] Algorithms, some researchers

[3] *See, e.g.*, Michal S. Gal & Niva Elkin-Koren, *Algorithmic Consumers*, 30 HARV. J.L. & TECH. 309, 318 (2017).

[4] *See, e.g.*, Anthony Sills, *ROSS and Watson Tackle the law*, IBM (Jan. 14, 2016), https://www.ibm.com/blogs/watson/2016/01/ross-and-watson-tackle-the-law [https://perma.cc/GA65-FDQD] (virtual attorneys can read and sort through more than a billion of documents per second and have the capacity to learn the law and get smarter over time); Amir Khandani et al., *Consumer Credit-Risk Models Via Machine-Learning Algorithms*, 34 J. BANKING & FIN. 2767 (2010) (algorithms used to determine credit risks).

[5] *See generally* ARIEL EZRACHI & MAURICE STUCKE, VIRTUAL COMPETITION (2016) (identifying four types of algorithmic conduct which can facilitate coordination); Salil K. Mehra, *Antitrust and the Robo-Seller: Competition in the Time of Algorithms*, 100 MINN. L. REV. 1323 (2016) (identifying the traits of algorithms which lead to coordination); Bruno Salcedo, Pricing Algorithms and Tacit Collusion 3 (Nov. 1, 2015) (unpublished manuscript) (on file with author) ("[W]hen firms compete via algorithms that are fixed in the short run but can be revised over time, collusion is not only possible but rather, it is *inevitable*." His results hold under specific assumptions regarding market conditions such as demand shocks that are more frequent than algorithm revisions.); *see generally* ORG. FOR ECON. CO-OPERATION & DEV., ALGORITHMS AND COLLUSION: COMPETITION POLICY IN THE DIGITAL AGE, 11–12 (2017) [hereinafter OECD]. For a more cautious view, see, e.g., Ulrich Schwalbe, Algorithms, Machine Learning, and Tacit Collusion 16 (Apr. 5, 2018) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3232631 [https://perma.cc/L3ZN-HEU7] ("[C]oordinated behaviour of algorithms is a possible outcome, but it is not as quick and easy or even unavoidable as it is often assumed."); Ashwin Ittoo & Nicolas Petit, Algorithmic Pricing Agents and Tacit Collusion: A Technological Perspective (Oct. 12, 2017)

argue, make coordination among suppliers easier and quicker than ever before. The higher levels of interconnection and transparency in digital markets, combined with more available data and a higher level of sophistication of analysis, makes reaching a joint profit-maximizing equilibrium easier. The speed and ease of detection and response to deviations from the coordinated equilibrium reduces incentives to break ranks. Joseph Harrington, Professor of Business Economics and Public Policy at Wharton Business School, argues that given developments in algorithmic agents, "the emergence of [coordination] . . . in actual market settings would seem extremely possible in the near future, if it is not already occurring."[6] Ariel Ezrachi and Maurice Stucke, Professors of Law at Oxford and the University of Tennessee, respectively, suggest in their seminal work on virtual competition that this effect is so strong, it marks the end of competition as we know it.[7]

Should algorithms indeed facilitate coordination in markets otherwise not prone to it, market participants and regulators need to explore what tools, if any, can be used to reduce the negative welfare effects of algorithmic-facilitated coordination on both consumer and social welfare.[8] While previous work suggested a (partial) market solution,[9] this Article focuses on legal remedies. In particular, this Article explores whether by applying laws that were designed to regulate human-facilitated market coordination we are limiting ourselves to looking only under the proverbial lamppost, while the activities we are interested in take place in the dark. If so, can we address this problem by using a stronger light bulb (i.e., widening the scope of existing laws by way of interpretation)? Or do we need to create a new source of light altogether (i.e., new laws)? Indeed, algorithms challenge the assumptions on which antitrust law is currently based. To illustrate, algorithms, unlike humans, can "read the minds" of other algorithms even before they perform any action, thereby transforming the need for an explicit commitment to coordinate or to

(unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3046405 [https://perma.cc/D6YV-J98R]. Coordination is not always welfare-reducing.

[6] Joseph E. Harrington Jr., Developing Competition Law for Collusion by Autonomous Price-Setting Agents 6 (Aug. 22, 2017) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3037818 [https://perma.cc/D8UP-Q7PM].

[7] *See* EZRACHI & STUCKE*, supra* note 5; *see also* LORDS SELECT COMMITTEE ON EUROPEAN UNION, ONLINE PLATFORMS AND THE DIGITAL SINGLE MARKET, REPORT, 2016-4, HL 129, ¶¶ 178–79 (UK) (acknowledging the rise of potential new means of collusion).

[8] For a short exposition, see Michal S. Gal, *Algorithmic-Facilitated Co-ordination: Market and Legal Solutions*, 1 COMPETITION POL'Y INT'L (2017).

[9] *See* Gal & Elkin-Koren, *supra* note 3, at 325–34.

punish deviations.[10] This new reality requires us to rethink concepts that stand at the basis of our laws, like the meeting of minds, intent, consent, and communication, and possibly requires us to create a new taxonomy to fit the algorithmic world. The analysis is timely: competition authorities all over the world are starting to explore such issues in depth, and the legality of algorithmic-facilitated coordination is likely to become a major issue, given rapid advancements in machine learning.

To meet this challenge, this Article advances in three interconnected stages (Part II–IV). Part II analyzes the effects of algorithms on the ability of competitors to coordinate their conduct. While this issue has been addressed by other researchers,[11] this Part of the Article seeks to contribute to the analysis by systematically charting the technological abilities of algorithms that may affect coordination in the digital ecosystem in which they operate. Part III explores the promises as well as the limits of market solutions. In particular, this Part considers the use of algorithms by consumers and off-the-grid transactions to counteract some of the effects of algorithmic-facilitated coordination by suppliers. The shortcomings of such market solutions lead to Part IV, which focuses on the ability of existing legal tools to deal effectively with algorithmic-facilitated coordination, while not harming the efficiencies they bring about. Further, this Article explores three interconnected questions that stand at the basis of designing a social welfare-enhancing policy: What exactly do we wish to prohibit, and can we spell this out clearly for market participants? What types of conduct are captured under the existing antitrust laws, thereby treating coordination-facilitating algorithms as illegal agreements? And is there justification for widening the regulatory net beyond its current prohibitions in light of the changing nature of the marketplace? The analysis refutes the Federal Trade Commission's acting Chairwoman's claim that current laws are sufficient to deal with algorithmic-facilitated coordination.[12]

[10] *See* John von Neumann, *First draft of a report on the EDVAC*, *in* 15 IEEE ANNALS HIST. COMPUTING 27, 33–34 (1993).

[11] Most notably by EZRACHI & STUCKE, *supra* note 5.

[12] Maureen K. Ohlhausen, Comm'r, Fed. Trade Comm'n, Remarks from the Concurrences Antitrust in the Financial Sector Conference: Should We Fear The Things That Go Beep in the Night? Some Initial Thoughts on the Intersection of Antitrust Law and Algorithmic Pricing (May 23, 2017) ("From an antitrust perspective, the expanding use of algorithms raises familiar issues that are well within the existing canon.").

II. ALGORITHMS AS COORDINATION FACILITATORS

Coordination among competitors is generally welfare-reducing: it lowers competitive pressures at the expense of price and choice.[13] Accordingly, the increased use of algorithms in the marketplace requires us to determine whether and to what extent algorithms facilitate coordination. To answer this, this Part first explores the conditions that must exist for coordination to take place; and then analyzes the ways that algorithms affect these conditions. As will be argued, while algorithms cannot facilitate coordination in all market settings, they can do so in a subset of markets, in which their characteristics enable competitors to overcome existing obstacles to coordination.

A. THE ECONOMICS OF COORDINATION

Competitors may have incentives to coordinate their conduct instead of competing among themselves. Nobel laureate economist, George Stigler, identified three cumulative conditions that must exist for such coordination to take place:[14]

1. *Reaching an understanding (or agreement)* on what trade conditions (e.g., price, quantity, or quality) will be profitable for all parties to the agreement. This means both resolving any disagreements as to the "correct" trade terms that all parties perceive as beneficial relative to a situation in which they do not coordinate, and communicating the ultimate decision to all parties.
2. *Detection of deviations* from the supra-competitive equilibrium. The slower and less completely deviations are detected, the weaker the coordination, as firms have stronger incentives to cheat. Also, if market conditions are not conducive to exposing deviations, firms seeking to detect deviations incur substantial costs. This reduces the overall attractiveness of coordination.
3. Creating a *credible threat of retaliation* in order to discourage deviations.

Economic theory further recognizes a fourth condition that must exist for coordination to take place:[15] *high entry barriers* in the market in which the

[13] *See, e.g.*, 11 PHILLIP E. AREEDA & HERBERT HOVENKAMP, ANTITRUST LAW: AN ANALYSIS OF ANTITRUST PRINCIPLES AND THEIR APPLICATION (1980) (suggesting exceptions exist when coordination is necessary to increase competition or efficiency).

[14] George J. Stigler, *Theory of Oligopoly*, 72 J. POLITICAL ECON. 44, 44–46 (1964).

[15] *See generally* ROBERT C. MARSHALL & LESLIE M. MARX, THE ECONOMICS OF COLLUSION (2012); Edward J. Green et al., *Tacit Collusion in Oligopoly*, *in* 2 OXFORD HANDBOOK OF INT'L ANTITRUST ECON. 464 (Roger D. Blair & D. Daniel Sokol eds., 2015). High entry barriers exist where the costs of new entry into a market are high.

coordinating parties operate. With low entry barriers, new competitors can easily enter and sweep the market, thereby reducing incentives to set supra-competitive trade terms in the first place.[16]

Economics and jurisprudence differ in their interpretations of Stigler's first condition: what constitutes reaching an agreement. In economic parlance, reaching an agreement captures both explicit agreements and conscious parallelism.[17] The former refers to cases where the parties exchange mutual assurances prior to their actions to act in a coordinated manner.[18] The latter, sometimes called oligopolistic coordination or tacit collusion, occurs when firms independently set their trade terms while taking into account their competitors' probable reactions to their actions.[19] In economic models, especially game theoretic ones, the specific method used to reach the agreement is not important.[20] However, as elaborated below, antitrust law is largely based on the distinction between these two situations. Only the former is considered to constitute "agreements" in the legal sense and is, therefore, potentially illegal; instances of conscious parallelism are not.[21]

The economics literature which deals with coordination among market players focuses on the market settings that must exist for Stigler's conditions to be fulfilled. As has been shown, even highly concentrated markets—in which only a small number of market players operate—can produce an uncertain market equilibrium, ranging from supra-competitive conditions, in which the trade terms offered to consumers are much less beneficial than under competitive conditions, to competitive ones.[22] Yet it is widely agreed that some market conditions and types of actions can make supra-competitive trade terms more likely, especially in a repeated market game.[23]

The economics literature identifies five broad categories of variables that affect Stigler's conditions: (1) market structure variables (e.g., market concentration, entry barriers), (2) product variables (e.g., product and cost homogeneity, multiplicity of products), (3) sales variables (e.g., secrecy), (4) demand variables (e.g., demand fluctuations, difficulties in estimating demand

[16] *See* MARSHALL & MARX, *supra* note 15.

[17] *See, e.g.*, William H. Page, *Tacit Agreement Under Section 1 of the Sherman Act*, 81 ANTITRUST L.J. 593, 593–94 (2017) (also noting that these terms have not been used consistently in case law or scholarly writings).

[18] *Id.* at 619.

[19] *Id.* at 601.

[20] *See* LOUIS KAPLOW, COMPETITION POLICY AND PRICE FIXING 8 (2013).

[21] *See* Page, *supra* note 17, at 602.

[22] JEAN TIROLE, THE THEORY OF INDUSTRIAL ORGANIZATION (1988).

[23] *See, e.g.*, Gregory J. Werden, *Economic Evidence on the Existence of Collusion: Reconciling Antitrust Law with Oligopoly Theory*, 71 ANTITRUST L.J. 719, 729–30 (2004).

for new products), and (5) the "personality" of the firms operating in the market (e.g., a tendency to act as a maverick).[24] The relevant factors may vary within a market over time, and some, such as entrepreneurial attitudes towards engagement in illegal activity, are intrinsically variable. Moreover, none of the factors are deterministic in their ability to facilitate coordination. Rather, they all reflect general tendencies subject to random deviations. In reality, a combination of market conditions will determine the likelihood of coordination. In what follows, I discuss some of the main coordination-facilitating factors.[25]

A concentrated market structure, where a small number of competitors are protected by high entry barriers, is a condition strongly conducive to coordination. This is because reaching an agreement to limit competition is easier and less costly if the number of firms involved is small.[26] With fewer firms to be checked for deviating conduct, detection of cheating is also easier. Furthermore, "[a] large number of firms not only makes it harder to identify a 'focal point' for co-ordination, but it also reduces the incentives for collusion as each player would receive a smaller share of the supra-competitive gains that an explicit or tacit collusive arrangement would be able to extract."[27]

Indeed, the number of firms is so important that it is largely assumed that conscious parallelism can only be reached in oligopoly markets (hence its alternative name, "oligopolistic coordination"). An oligopoly exists when a small number of firms dominate the market.[28] The main economic characteristic of oligopolistic markets is that each firm's decisions have a noticeable impact on the market and on its competitors.[29] Though each firm may strategize independently, any rational decision must take into account the anticipated reaction to its decisions by competitor firms.[30] The decisions of firms in an oligopoly may thus be interdependent even though arrived at independently. Such mutual interdependence may forestall competitive conduct.

Transparency of transactions makes it easier to coordinate market offers, to detect deviations, and to determine the level of sanctions that should be

[24] *See, e.g.*, TIROLE, *supra* note 22.

[25] *See, e.g.*, MARC IVALDI, BRUNO JULLIEN, PATRICK REY, PAUL SEABRIGHT & JEAN TIROLE, THE ECONOMICS OF TACIT COLLUSION 11 (2003); *see generally* SIGRID STROUX, US AND EC OLIGOPOLY CONTROL (2004).

[26] IVALDI ET AL., *supra* note 25, at 12.

[27] OECD, *supra* note 5, at 20–21.

[28] Carl Shapiro, *Theories of Oligopoly Behavior*, *in* HANDBOOK OF INDUSTRIAL ORGANIZATION 329 (R. Schmalensee & R.D. Willig eds., 1st ed. 1989).

[29] *Id.*

[30] *Id.*

applied to deviators.[31] Furthermore, transparency in any firm's decisional parameters and in the inputs used in the decision making process make it simpler for others to understand what is driving their competitors' actions.[32] As a result, this makes it easier to reach an agreement and limits the instances in which a mistaken categorization of a competitor's actions could lead to a price war.[33]

The availability of information also affects coordination: the noisier or more incomplete the information, the harder it is to coordinate.[34] Along those lines, demand fluctuations make it more difficult to set a stable, jointly profitable price. They also make detection of deviations much harder and increase the chance of a price war.[35] Consider the following example: a supplier observes that demand for his product is reduced. He cannot effectively differentiate between natural changes in consumer demand, which are likely to affect all suppliers in the market (or even mainly his product if products are heterogeneous), and deviation from the status quo on the part of a competing supplier who now enjoys a larger market share. Both possibilities may lead the supplier to lower his prices, potentially triggering a price war. It may take time until coordination is once again achieved, if at all. Accordingly, the more imperfect the price signals among suppliers, the less stable the coordination.

Economic studies have also shown that pre-play communication among suppliers is important for coordination.[36] Indeed, experiments on oligopolies have shown that absent communication, tacit collusion is not easy to achieve.[37] Cooper and Kuhn show that explicit threats to punish cheating are the most important factor in successfully establishing coordination, once a cooperative strategy is established.[38]

Where market conditions create obstacles to coordination, firms may take more direct actions that facilitate coordination (or purposefully refrain from certain actions that limit it). Such actions include behavior that helps firms

[31] IVALDI ET AL., *supra* note 25, at 25.

[32] *Id.* at 26.

[33] *Id.* at 25–26.

[34] *See* Schwalbe, *supra* note 5, at 12.

[35] Edward J. Green & Robert H. Porter, *Noncooperative Collusion under Imperfect Price Information*, 52 ECONOMETRICA 87, 94–95 (1984).

[36] *See, e.g.*, Joseph E. Harrington, Jr., *How do Cartels Operate?*, 2 FOUND. & TRENDS IN MICROECONOMICS 1 (2006); Yu Awaya & Vijay Krishna, *On Communication and Collusion*, 106 AM. ECON. REV. 285 (2015).

[37] *See, e.g.*, Jan Potters & Sigrid Suetens, *Oligopoly Experiments in the Current Millennium*, 27 J. ECON. SURVEYS 439 (2013); Niklas Horstmann, Jan Krämer, & Daniel Schnurr, *Number Effects and Tacit Collusion in Experimental Oligopolies*, J. INDUS. ECON. (forthcoming).

[38] David J. Cooper & Kai-Uwe Kühn, *Communication, Renegotiation, and the Scope for Collusion*, 6 AM. ECON. J. MICROECONOMICS 247, 268 (2014).

overcome the complicating factors that make coordination infeasible or insufficient to yield monopoly profits.[39] Such practices may range widely, from standardizing products or notifying competitors of upcoming changes in prices, to signaling how one will react to market changes.[40] They can be adopted either by agreement or unilaterally.[41] Accordingly, both the market's natural conditions, as well as actions taken by market players, affect the ability to meet Stigler's three conditions for coordination.

B. How Algorithms Facilitate Coordination

Can algorithms affect the market equilibrium and facilitate coordination? To answer this question, we need to explore how algorithms may affect the conditions for coordination explored above.

Addressing this issue requires us to combine insights from computer science and economics. Computer science brings light on the technological side as to how algorithms operate, and their comparative advantages and limitations. Economics brings light on the market equilibria that will most likely ensue, given the market conditions created by algorithms. Below, I briefly explore insights from both disciplines. I start by briefly reviewing the characteristics of algorithms, and then relating them to the ability to facilitate coordination.

1. *What Are Algorithms?*

Algorithms are structured decision-making processes that automate computational procedures to generate decisional outcomes on the basis of data inputs.[42] In a broad sense, we all use (non-automated) algorithms in our daily lives. For example, when we decide what to wear, we use data inputs (such as the weather, the occasion, and comfort) and weigh them in order to reach an outcome that most accords with our preferences (e.g., one cannot wear a comfortable jumpsuit to a formal party). Coded algorithms do the same. They use predetermined decision procedures in order to suggest a decision, given particular data.[43]

[39] *See* George A. Hay (1984), *Facilitating Practices: The Ethyl Case*, *in* The Antitrust Revolution: Economics, Competition, and Policy 182, 189 (Kwoka & White eds., 3rd ed. 1999).

[40] *See, e.g.*, Steven C. Salop, *Practices that (Credibly) Facilitate Oligopoly Coordination*, *in* New Developments in the Analysis of Market Structure 265, 271 (Joseph E. Stiglitz & G. Frank Mathewson eds., 1986); William H. Page, *Facilitating Practices and Concerted Action under Section 1 of the Sherman Act*, *in* Antitrust Law and Economics 23 (Hylton ed., 2010).

[41] *See* Salop, *supra* note 40.

[42] *See* Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest & Clifford Stein, Introduction to Algorithms 5 (3rd ed. 2009).

[43] *Id.* at 192–93, 843–49.

Algorithms vary significantly in the computational procedures they use (such as sorting or merging data, finding correlations, etc.) and their efficiency in achieving the given task (including time, amount of data, and computer power needed to complete a task).[44] Importantly for our analysis, algorithms can operate at different levels of abstraction. At the lowest level, all parameters are dictated by the developer in advance ("expert algorithms").[45] Such pre-selection of relevant features enables the algorithm to operate more quickly, and also reduces the amount of data needed.[46] Yet such pre-selection is rigid in the sense that if correlations in the data change over time, the algorithmic decision will not reflect this. Accordingly, algorithms can be designed to set or to refine their own decision parameters in accordance with the data inputted in them and the decision-making techniques they are coded to perform ("learning algorithms").[47] Learning algorithms employ machine learning—a type of artificial intelligence that gives computers the ability to learn from the data they encounter without the need to define correlations a priori.[48] Accordingly, learning algorithms do not follow strictly static program instructions, but rather build a decision process by learning from data inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or unfeasible (common examples include spam filtering and optical character recognition).[49] While machine learning identifies correlations between data inputs, it usually does not explain the causality of such correlations.[50] Some algorithms combine the functions of expert and learning algorithms.[51]

In today's world, characterized by big data, fast digital connectivity, and increased computational and storage capacity, algorithms may create significant advantages in decision-making. The most basic advantage they offer

[44] *Id.* at 5–6.

[45] OECD, *supra* note 5, at 11–12.

[46] *See* Yann LeCun, Yoshua Bengio & Geoffrey Hinton, *Deep Learning*, 521 NATURE 436, 436 (2015).

[47] *See, e.g.*, OECD, *supra* note 5, at 9–11. For examples of machine learning already used in algorithms, see Ariel Ezrachi & Maurice E. Stucke, *Artificial Intelligence & Collusion: When Computers Inhibit Competition*, 2017 U. ILL. L. REV. 1775 (2017).

[48] *See generally* TOM MITCHELL, MACHINE LEARNING (1997). Other types of artificial intelligence include, for example, expert systems, which use databases of expert knowledge, to offer advice on make decisions in such as areas as medical diagnosis of stock exchange trading.

[49] OECD, *supra* note 5, at 11–13.

[50] Some advanced algorithms can also find causality. *See, e.g.*, Rainer Opgen-Rhein & Korbinian Strimmer, *From Correlation to Causation Networks: A Simple Approximate Learning Algorithm and Its Application to High-Dimensional Plant Gene Expression Data*, 1 BMC SYSTEMS BIOLOGY 37 (2007).

[51] Schwalbe, *supra* note 5, at 15.

is speed in the collection, organization, and analysis of data, enabling exponentially quicker decisions and reactions.[52] The vast volume of data now available, which challenges the human cognitive capacity to process the relevant information, has made this ability even more important.[53] Given any number of decisional parameters and data sources, computers can generally apply the relevant algorithm at a velocity unreachable by the human brain, especially if the decision involves a large number of parameters that need to be balanced or many data inputs that must be analyzed or compared.[54] Automatic acceptance of the algorithm's suggestion further enables an exponentially quicker reaction. As innovator Elon Musk observed, "[a] computer can communicate at a trillion bits per second, but your thumb can maybe do . . . 10 bits per second or 100 if you're being generous."[55]

The second main advantage of algorithms relates to their analytical sophistication. Advances in data science, including data collection and storage, have ushered in the age of big data, which enables algorithms to integrate numerous variables into their decisions.[56] This provides a level of sophistication that cannot be achieved by the human mind without substantial time and effort. In one noteworthy example, algorithms defeated world champions in the strategic game Go.[57]

It is thus not surprising that the use of algorithms to make commercial decisions is spreading fast. Algorithms are used in a myriad of tasks, including responding rapidly to changes in demand conditions, determining efficient levels and locations for production and storage, and assessing risk levels.[58] Important for our analysis, they are also used for pricing decisions.[59] Some common examples include Uber's surge pricing algorithm, which is used to set

[52] *See, e.g.*, OECD, *supra* note 5, at 15; Harrington, *supra* note 6, at 54. For an example, see the velocity of facial recognition though an algorithm: PATRICK GROTHER, MEI NGAN, & KAYEE HANAOKA, ONGOING FACE RECOGNITION VENDOR TEST (FRVT) (2018).

[53] For the importance of data, see, e.g., Avigdor Gal, It's a Feature, Not a Bug: On Learning Algorithms and What They Teach Us (unpublished Note for the 127th meeting of OECD Roundtable on Algorithms and Collusion 21–23 June 2017).

[54] Harrington, *supra* note 6.

[55] Steve Renick, *Elon Musk at the World Government Summit 2017 in Dubai. Conversation with Mohammad AlGerga*, YOUTUBE (June 22, 2017), https://www.youtube.com/watch?v=R5dHlLjOdjk [https://perma.cc/G4RQ-SJJN].

[56] *See, e.g.*, Matthew Adam Bruckner, *The Promise and Perils of Algorithmic Lenders' Use of Big Data*, 93 CHI.-KENT L. REV. 3 (2018).

[57] *See* Paul Mozur, *Google's AlphaGo Defeats Chinese Go Master in Win for A.I.*, N. Y. TIMES (May 23, 2017), https://www.nytimes.com/2017/05/23/business/google-deepmind-alphago-go-champion-defeat.html [https://perma.cc/S8FU-4PPQ].

[58] *See generally* OECD, *supra* note 5; Rob Kitchin, *Thinking Critically About and Researching Algorithms*, 20 INFO., COMMC'N & SOC'Y 14 (2017).

[59] *See* OECD, *supra* note 5, at 16; *see also* Ezrachi & Stucke, *supra* note 5.

prices based on demand and supply conditions, and the algorithm used by Airbnb to price differentiated offers.[60] In parallel to the creation of tailor-made algorithms, software firms also sell off-the-shelf pricing algorithms, which can be relatively easily fit to each supplier's needs.[61] Some examples include Feedvisor's self-learning algorithmic repricer, which uses artificial intelligence and big data techniques to set prices,[62] or Inoptimizer, a pricing engine based on artificial intelligence and data on competitors' and consumers' behavior.[63] Sophisticated algorithms often treat pricing as a reinforcement learning issue, changing their decision matrix in an ongoing way as they learn from market interactions.[64]

Algorithms can also be used to learn how other business entities set their trade conditions. They can do this by directly observing and analyzing the code of other algorithms, or by analyzing competitors' behavior under given market conditions to indirectly learn their decisional parameters.[65] Algorithms can also police other firms, by determining when another firm has strayed from the status quo and by setting trade conditions designed to deter firms from doing so.[66]

The ability of algorithms to achieve their function is contingent on several factors. The first is the quality and volume of the data used by the algorithm as inputs. The best theoretical model will only work well if it has the necessary information on which to base its decisions.[67] Accordingly, the ability of firms to access data which is necessary in order to determine the coordinated outcome affects their ability to coordinate. Data can come from many sources, including the Internet, sensors placed in physical goods ("Internet of Things"),

[60] *See, e.g.*, Ryan Calo & Alex Rosenblat, *The Taking Economy: Uber, Information, and Power*, 117 COLUM. L. REV. 1623, 1656 (2017); Kenneth A. Bamberger & Orly Lobel, *Platform Market Power*, 32 BERKELEY TECH. L.J. 1051, 1071 (2017); Shagun Jhaver, Yoni Karpfen & Judd Antin, *Algorithmic Anxiety and Coping Strategies of Airbnb Hosts*, *in* PROCEEDINGS OF THE 2018 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (2018).

[61] *See, e.g.*, OXERA, WHEN ALGORITHMS SET PRICES: WINNERS AND LOSERS (2017).

[62] *Amazon Algorithmic Repricer*, FEEDVISOR, https://feedvisor.com/amazon-repricer/ [https://perma.cc/T4VV-7KB9] (last visited Mar. 18, 2019).

[63] *Inoptimizer*, INTELLIGENCE NODE, http://www.intelligencenode.com/products-inoptimizer.php [https://perma.cc/QBY2-KJ2Q] (last visited Mar. 18, 2019).

[64] *See generally* RICHARD S. SUTTON & ANDREW G. BARTO, REINFORCEMENT LEARNING: AN INTRODUCTION (2017).

[65] Salcedo, *supra* note 5, at 2, 8–10.

[66] Ariel Ezrachi & Maurice E. Stucke, Algorithmic Collusion: Problems and Counter-Measures 4, 10 (unpublished manuscript for the 127th meeting of OECD Roundtable on Algorithms and Collusion 21–23 June 2017).

[67] *See* Chris Brummer & Yesha Yadav, *Fintech and the Innovation Trilemma*, 107 GEO. L.J. 235, 276 (2019).

and human interviews.[68] It can also often be bought on the market as a commodity.[69] The more accurate the data, and the faster it can be analyzed, the stronger the ability to coordinate.

Performance is also affected by the quality and speed of the data analysis performed by the algorithm. A sophisticated or efficient algorithm might be able to mine the needed information from lower-quality data.[70] The computer's computational power and its ability to store and quickly retrieve data also affect performance. Finally, the computational procedure used by the algorithm affects performance. To illustrate, compare two paradigmatic cases: In the first one, algorithms react only to changes in input prices. In the second, algorithms react to changes in input prices and to prices set by competitors. Clearly, the second algorithm is more conducive to coordination.

2. Can Algorithms Affect Coordination?

Let us now relate the characteristics of algorithms to their ability to facilitate coordination. Although economists have yet to study in-depth the effects of algorithms on coordination, researchers are already split in their views of whether algorithms make a difference. While most researchers argue that at least under some market conditions, algorithms can make coordination more likely, others are more cautious, especially with regard to the design of autonomous algorithms that operate in complex settings.[71] Notably, most studies do not analyze the effect on coordination of the characteristics of algorithms and of the digital world in which they operate in a systematic manner.[72] This Article attempts to contribute to this important debate by doing so.

The analysis below assumes that the fourth condition for coordination—the existence of high entry barriers—is fulfilled. In markets where this is not true, a supra-competitive price will not be sustainable. Does the use of algorithms itself heighten entry barriers? Not necessarily, though in certain circumstances, in which the algorithm's special qualities or the unique dataset

[68] *See, e.g.*, MAURICE STUCKE & ALLEN GRUNES, BIG DATA AND COMPETITION Policy (2016); JAMES MANYIKA ET AL., BIG DATA: THE NEXT FRONTIER FOR INNOVATION, COMPETITION, AND PRODUCTIVITY 21–22 (2011).

[69] *See generally* Herbert Zech, Data as a Tradeable Commodity – Implications for Contract Law 1 (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3063153 [https://perma.cc/25MS-U4DG].

[70] *See, e.g.*, Brummer & Yadav, *supra* note 67.

[71] *See supra* note 3. For more cautious views on the ability of algorithms to coordinate see Ittoo & Petit, *supra* note 5.

[72] For an exception, see, e.g., Ariel Ezrachi & Maurice E. Stucke, *Sustainable and Unchallenged Algorithmic Tacit Collusion* (Univ. of Tenn. Legal Studies Research Paper No. 266, Dec. 6, 2018).

on which it operates cannot be copied or easily reconstructed[73] (e.g., Google's database), the algorithm (or the data used in it) may create a significant comparative advantage.[74] Regardless, this Article focuses on cases in which entry barriers—of any origin[75]—are presumed to be high.

Where entry barriers are high, I argue that reaching a supra-competitive equilibrium by using algorithms operating in our digital world can be easier, relative to a similar market operating without algorithms. To show this, I explore how algorithms affect Stigler's three conditions.

Stigler's first condition, reaching an agreement (in economic parlance), is made easier by the use of algorithms. Several factors combine to reduce the difficulty in calculating a joint profit-maximizing equilibrium: (1) the greater availability of data, particularly real-time and more accurate data on market conditions, including digital price offers of competitors and suppliers of intermediate goods and services, as well as data on consumer preferences; (2) cheaper and easier data collection and storage tools (e.g., the cloud);[76] (3) advances in Internet connectivity which allow for cheaper and faster transfer of data;[77] and (4) the increasingly strong and sophisticated analytical power of algorithms due to advances in data science.[78]

Indeed, algorithmic sophistication makes it easier to solve the multidimensional problems raised by coordination, such as establishing a jointly profitable price in a market with differentiated products. Algorithms can be used not only to perform a single action, but also to determine and execute complex contingent strategies. Algorithmic sophistication also implies that fewer repeated games might be needed to reach a coordinated equilibrium. Indeed, studies performed by Google's artificial intelligence business, DeepMind, on algorithmic interactions found that algorithms with more

[73] *See, e.g.*, Daniel L. Rubinfeld & Michal S. Gal, *Access Barriers to Big Data*, 59 ARIZ. L. REV. 339, 373 (2016).

[74] *Id.* at 354.

[75] Some conditions which characterize the digital world affect the height of entry barriers. For example, increased connectivity between consumers and suppliers through the Internet reduces the need to open physical stores. *See, e.g.*, Gal & Elkin-Koren, *supra* note 3, at 329. Yet large digital platforms that connect consumers and suppliers may provide the platform owner with advantages in data collection, and so may increase entry barriers. STUCKE & GRUNES, *supra* note 67.

[76] Availability of data depends on the height of entry barriers into big data markets. *See generally* Rubinfeld & Gal, *supra* note 73.

[77] In an EU study, approximately half the retailers who answered the questionnaire said they track online prices, and most use automatic software programs, sometimes called crawlers. *See Final Report on the E-commerce Sector Inquiry*, at 51, COM (2017) 229 final (May 10, 2017).

[78] *See* discussion *infra*.

cognitive capacity sustained more complex cooperative equilibria.[79] Yet in situations in which the complexity of cooperation was too high or it was not rational to cooperate, the algorithms competed vigorously.[80] This implies that algorithms are subject to limitations, even if these are less demanding than those faced by humans performing similar tasks. Given the high stakes involved and the pace of technological developments in machine learning, it is envisioned that at least some of these technological limitations will be alleviated.[81]

Machine learning has the potential to play an important part in reaching a coordinated outcome. The algorithm may learn, even before it starts to operate in the market, when and which coordination is optimal. Such learning can be supervised or unsupervised. Supervised learning involves a process in which the algorithm determines the decisional parameters through an externally supervised training process, in which it is corrected when its predictions are incorrect.[82] The training process continues until the algorithm achieves a desired level of accuracy. Unsupervised learning involves a process in which the algorithm autonomously determines the decisional parameters by deducing decisional rules from correlations found in the input data (such as how past pricing patterns affected profitability).[83] Machine learning may thus enable the algorithm identify the best reactions to market conditions, given specified data. The artificial intelligence literature, while focusing on social dilemmas rather than on pricing issues, has shown that learning can lead to cooperative outcomes.[84]

Observe that to be jointly profitable, the coordinated price need not be the perfect profit-maximizing price (i.e., the Pareto optimal one, which is the highest price which still maximizes the firms' profits). For that to happen, firms may need data on factors such as the real production costs and production capacities of their competitors.[85] In some situations, such information can be indirectly observed or calculated, even if not perfectly. In

[79] Leibo et al., *Multi-agent Reinforcement Learning in Sequential Social Dilemmas*, *in* PROCEEDINGS OF THE 16TH INTERNATIONAL CONFERENCE ON AUTONOMOUS AGENTS AND MULTIAGENT SYSTEMS 464, 471 (2017); *see generally* Ittoo & Petit, *supra* note 5, at 10–13.

[80] *See* Leibo et al., *supra* note 79, at 467.

[81] Ittoo & Petit, *supra* note 5, at 13.

[82] *See, e.g.*, Schwalbe, *supra* note 5, at 8.

[83] *See, e.g.*, *id.* at 9.

[84] *See, e.g.*, Dipyaman Banerjee & Sandip Sen, *Reaching Pareto-Optimality in Prisoner's Dilemma Using Conditional Joint Action Learning*, 15 AUTONOMOUS AGENT & MULTI-AGENT SYSTEMS 91 (2007); Leibo et al., *supra* note 79.

[85] *See, e.g.*, Susan Athey & Kyle Bagwell, *Collusion with Persistent Cost Shocks*, 76 ECONOMETRICA 493 (2008).

a repeated game, firms can signal such factors to each other, or the algorithm might be based on a profit-maximizing benchmark that was previously used in that market. Yet even when such information is not completely observable, firms may still find it profitable to coordinate so long as the price is the best approximation of the maximal price that can be set with the existing data, and is greater than the price which would have been set absent coordination. Hence, the fact that algorithms may not reach the perfect equilibrium does not lead to the conclusion that algorithms cannot facilitate coordination.

The fact that algorithms—unless their developers code them otherwise—make rational decisions, devoid of ego and biases, also potentially eases coordination, by making their decisions more predictable.[86] However, this factor could also lead in the other direction. "Rational" algorithms may be less affected than humans by forces such as guilt aversion, lying aversion, and group identity, which increase adherence to agreements and leads to more stable cooperation.[87] Much depends, of course, on the extent to which market players treat defection by an algorithm differently from defection by a human being.

A third effect of algorithms, which promotes Stigler's first condition, is that they shorten time lags of reaching new equilibriums when market conditions change. The speed and sophistication of algorithms, combined with the increased availability of real-time data and faster connectivity, enable them to quickly recognize changes in market conditions and to autonomously change their decisional parameters accordingly.[88] As a result, a new agreement is much easier and quicker to reach.

Fourth, and importantly for the legal analysis that follows, algorithms change the mode and dynamics of communication needed to reach an agreement. As John von Neumann, one of the founding figures of computer science, observed more than half a century ago, algorithms serve a dual purpose: as a set of instructions, and as a file, to be read by other programs.[89] The first use relates to the fact that an algorithm is a pre-set decision

[86] *See generally* Ezrachi & Stucke, *supra* note 47, at 1792; Jan Blockx, *Antitrust in digital markets in the EU: policing price bots*, *in* DIGITAL MARKETS IN THE EU 75 (J. M. Veenbrink, ed., 2018). Observe that biases can nonetheless arise from biased data which is inputted into the algorithm.

[87] *See, e.g.*, Robyn M. Dawes, Jeanne McTavish & Harriet Shaklee, *Behavior, Communication, and Assumptions About Other People's Behavior in a Commons Dilemma Situation*, 35 J. PERSONALITY & SOC. PSYCHOL. 1 (1977); Gary Charness & Martin Dufwenberg, *Promises and Partnership*, 74 ECONOMETRICA 1579 (2006).

[88] For the ability of algorithms to change the decision parameters autonomously, see, e.g., Schwalbe, *supra* note 5, at 9.

[89] Neumann, *supra* note 10, at 1–2.

mechanism, a "recipe" for making decisions.[90] The second use relates to the fact that algorithms can be instructed to read other algorithms, and to perform some action if the other program's content is of a particular kind.[91] This simple but fundamental idea highlights a central difference between human and algorithmic coordination: when an algorithm is transparent to others, another algorithm can "read its mind" and accurately predict all its future actions when given any specific sets of inputs, including changes in market conditions and reactions to other player's actions. Indeed, as Moshe Tennenholtz, Professor of Computer Science at the Technion has proven, this unique characteristic means that coordination can often be achieved in a one-shot game.[92] This is not true with regard to human interaction, in which one cannot accurately "read the mind" of another and predict all future actions. This algorithmic trait can also serve to limit misguided price wars.

To make this fundamental change in communication methods clearer, let us use a simple example. Player A adopts the following algorithm:

Algorithm A:

```
Calculate best joint price under assumption that my price=Price set by Algorithm B;
Set my price accordingly;
Wait 10 seconds;
Search for price set by algorithm B;
If price set by algorithm B (larger or equal to) my price then repeat this set of actions every 5 seconds (loop);
Else reduce my price by 50%.
```

Player B reads and understands the decision process adopted in Algorithm A, which enables it to accurately predict player A's reactions to changes in market conditions and to his prices. Algorithm A serves both as a self-commitment device, an indication of course for future action, and as an explicit threat of retaliation. B will then have strong incentives to adopt the following algorithm, should the price set by A be sufficiently close to the jointly profitable price:

[90] *Id.*

[91] Even if different computer languages are used, an algorithm can "translate" the code.

[92] Moshe Tennenholtz, *Program Equilibrium*, 49 GAMES & ECON. BEHAV. 363, 364 (2004).

Algorithm B:

```
Search for price set by algorithm A;
Set my price=price set by algorithm A;
Repeat this set of actions every 5 seconds (loop).
```

Algorithm B instructs the computer to compare player B's price to that of player A. This decision parameter is a rational reaction to the "price recipe" of Algorithm A. It also serves to motivate player A not to deviate, because any lower price he sets will be matched by B. This motivation is strengthened by the speed at which monitoring and reactions (price changes) take place. Indeed, the interaction between the players is based on each reasoning computationally about the other's algorithm.

The result is coordinated pricing as a direct consequence of simple leader-follower behavior, where B acts solely based on information about A's prices, which are available online. Moreover, although the interaction is asynchronous (since each reacts to prices set by the other), the speed of the Internet makes the resulting price changes almost synchronous.[93]

As the above example indicates, the use of an algorithm can send a strong and clear signal to other market players about several factors that are important for coordination:

1. The decisional parameters on which the algorithm will set its price, which can be observed by other market players even before any action is actually taken (A: Calculate best joint price under assumption that my price=Price B; B: Set my price=Price A);
2. The frequency of searches for deviations (A: Wait 10 seconds, and search for Price B; B: Repeat every 5 seconds);
3. The punishment for deviation by switching from a high payoff to a low payoff continuation equilibrium (A: Otherwise reduce my price by 50%; B: [Always] Set my-price=Price A).

Accordingly, this recipe for action, which contains an entire contingent plan for coordination in a few lines of code, creates both pre-agreement communication that the other party can "read" and understand, and a self-commitment device. It also increases the level of certainty for both parties. For

[93] This example applies where both suppliers sell homogenous goods. However, as the Topkins case suggests, a more sophisticated algorithm can be used to set jointly profitable prices in much more complicated settings. There, the sellers sold different posters, in infrequent transactions. *See* Topkins, *infra* note 191.

example, both players are certain about what punishment to expect. Importantly, the use of algorithms limits the need for some forms of communication (e.g., verbal assurances of commitment or advance price change announcements) that were seen as necessary for establishing cooperation in a world based on human coordination.[94]

This implies that communication to competitors of future intended actions can be performed by simply making one's algorithm transparent and readable by (select) others' communication protocols.[95] The fact that the information is observable online eliminates the need to "drive by your competitor's petrol station" to know what he will charge, and creates immediate visibility of one's trade terms to multiple competitors. Moreover, to achieve transparency, the algorithm need not be directly observable. As the economist Bruno Salcedo argues, the analytical qualities of algorithms can be utilized to determine the decision processes of other algorithms, provided that the former have sufficient information about the decisions made by the latter under changing market conditions.[96] While it is more difficult to create transparency where decisions are taken by algorithms based on neural networks in which the decision process is not easily observable or explainable, if the specific neural network is transparent and can be copied, or if correlations in the algorithm's data inputs and outputs are observable, then the algorithm's outcomes may be predictable.

This observation cannot be overstated: the mere (direct or indirect) observation of the algorithm by competitors may, by itself, serve to facilitate coordination. As economic studies show, the ability to communicate price choices in oligopolistic markets may drastically change the market equilibrium, as collusion increases substantially and significantly.[97] The algorithm can

[94] William E. Kovacic, Robert C. Marshall, Leslie M. Marx & Halbert L. White, Jr., *Plus Factors and Agreement in Antitrust Law*, 110 MICH. L. REV. 393, 417 (2011).

[95] Recent studies focus on how machine learning can be used to let algorithms automatically discover and create the communication protocols needed to coordinate their behavior. Some examples include Sainbayar Sukhbaatar, Arthur Szlam & Rob Fergus, *Learning Multiagent Communication with Backpropagation*, 29 ADVANCES NEURAL INFO. PROCESSING SYS. 2252 (2016) (demonstrating the ability of algorithms to learn to communicate among themselves by creating a communication protocol); Jakob N. Foerster et al., Learning to Communicate to Solve Riddles with Deep Distributed Recurrent Q-Networks (2016) (unpublished manuscripts), https://arxiv.org/pdf/1602.02672.pdf [https://perma.cc/Z7F8-UJ2W] (creation of communication among algorithms for tasks which are fully cooperative, partially observable, sequential multi-agent decision-making problems. Communication is learned and agents communicate through actions). While these abilities might not, as of yet, be applied to pricing algorithms, they are likely to be added given rapid progress in artificial intelligence.

[96] Salcedo, *supra* note 5.

[97] *See* Christoph Engel, *Tacit Collusion: the Neglected Experimental Evidence*, 12 J. EMPIRICAL LEGAL STUD. 537 (2015).

communicate much more than price choices: it communicates a business strategy. Such communications need not be binding, but algorithms may strengthen this aspect as well.

This raises the question of the motivation of the user to make its algorithm and the data which it uses transparent. Exclusive access to algorithms and data can create a comparative advantage, and thus may be regarded as important trade secrets not to be shared with others. Yet at least some factors favor an inclination toward transparency. First, an important difference exists between firms whose comparative advantage lies in the creation of a pricing algorithm, and those in which it lies elsewhere. The latter have weaker incentives to protect the secrecy of their pricing algorithms and the data they rely on. Second, encoding can be used to create selective transparency. Third, transparency need only relate to the pricing part of the algorithm and not to all its functions. Finally, motivations for transparency will be determined by the balance between the increased profitability from coordination relative to the profitability of operating without it.[98]

The foregoing analysis also suggests that Stigler's second condition, detection of deviations from the status quo, is fulfilled more easily and quickly by algorithms. Due to their high levels of sophistication and reduced ingrained biases, algorithms may better differentiate between intentional deviations from coordination and natural reactions to changes in market conditions or even errors, which change the efficient status-quo, thereby preventing unnecessary price wars.[99]

Interestingly, the incentives to deviate in the first place are also reduced. Since technology enables the algorithm to react almost immediately to changes in a competitor's price, consumers may not be aware of ephemeral price differences between competitors and therefore may not switch between them. Competitors, acknowledging this fact, have weaker incentives to deviate.[100] Furthermore, the fact that digital markets have made it much easier for consumers to conduct transactions themselves has increased the number of small and frequent purchases. This, in turn, further reduces incentives to deviate, since the benefits from deviation are likely to be small and temporary. By way of analogy, correcting a mistaken assumption in an algorithm-driven market is like correcting a wrong turn on a road with many intersections, as opposed to accidentally getting on a highway with long stretches between interchanges. Thus, this reduces the need for credible punishments that devalue extra profits made during the deviation period, while at the same time

[98] EZRACHI & STUCKE, *supra* note 5.

[99] OECD, *supra* note 5, at 22.

[100] *Id.*

increasing the credibility of an immediate switch from a collusive equilibrium to a competitive one, which would lower profits for all players in the future. In such an environment, changes in price may be almost immediately rescinded.

Stigler's third condition, creating a credible and sufficiently strong threat of retaliation against deviators, can also be more easily met by algorithms. Given their potentially high level of sophistication, algorithms can better calculate the level of sanctions necessary to discourage deviations. Moreover, algorithms may create a credible threat of retaliation, if changing their decision tree is not simple or change may take a long time relative to the frequency of market transactions.

Based on the foregoing analysis, algorithms operating in digital markets may facilitate coordination in three ways. First, they ease the fulfilment of Stigler's conditions. Second, and more interestingly, algorithms lessen the need to commit to Stigler's conditions a priori. As elaborated, they can more quickly recalculate one's optimal reaction, thereby reducing the need for an optimal equilibrium in the first round, and they lower incentives to deviate, thereby reducing the need for explicit ex ante commitments or threats of strong punishment.[101] Accordingly, algorithms operating in the digital world increase the likelihood of coordination without the need for strong pre-action commitments and threats. Finally, algorithms may strengthen not only players' ability to reach an agreement, but also their incentives to do so. One factor which affects such incentives is the risk of detection by enforcement agencies and private plaintiffs. A study performed by Google Brain has shown that algorithms can autonomously learn how and when to encrypt messages, given a specified secrecy policy, in order to exclude other algorithms from the communication.[102] Unless third parties have a way of determining when the conduct of algorithms is based on such encryption, detection will become much harder. Furthermore, should algorithmic signaling and interactions be sufficient to sustain a supra-competitive equilibrium, algorithms reduce the need to meet in the real world, thereby further reducing the chances of getting caught. Accordingly, in markets where entry barriers are high and algorithms can facilitate meeting the conditions for coordination, the appearance and stability of supra-competitive prices may increase.[103]

[101] *See supra* Section II.B.1

[102] *See generally* Martin Abadi & David G. Anderson, Learning To Protect Communications with Adversarial Neural Cryptography (2016) (unpublished manuscript), https://arxiv.org/pdf/1610.06918v1.pdf [https://perma.cc/AG5C-73UQ].

[103] *See also* OECD, *supra* note 5, at 35 ("Algorithms might affect some characteristics of digital markets to such an extent that tacit collusion could become sustainable in a wider range

This is not to say that algorithms can facilitate coordination in all circumstances. Where entry barriers are low, or where one or more of Stigler's conditions cannot be effectively met, coordination will not take place. This may be the case, for example, in markets where demand fluctuations are significant and difficult to distinguish from deviations from the equilibrium, or where the relevant data are not easily accessed by all competitors.[104] As Ashwin Ittoo and Nicolas Petit, Professors of Information Systems and Law, respectively, at the University of Liege, argue,"[w]hile we do not deny the fact that smart pricing agents can enter into tacit collusion and that regulators may be right to be vigilant, we find that there are several technological challenges in the general realm of [reinforcement learning] that mitigate this risk."[105] In particular, current algorithmic sophistication may not be sufficient to overcome coordination obstacles in complex setting, especially where competitors lack information on their rivals' business strategies, input prices, and demand forecasts.[106] Indeed, algorithms provide no panacea to these coordination problems, which similarly plague human-facilitated coordination. Nonetheless, as shown above, at least in some circumstances, algorithms may be able to reduce their significance. For example, business strategies can be communicated though the coding and transparency of the algorithms.[107] Furthermore, given the high profits to be had from coordination, it is envisioned that computational complexity problems[108] will be reduced as firms develop more sophisticated algorithms.[109]

of circumstances possibly expanding the oligopoly problem to non-oligopolistic market structures.").

[104] *See* Edward J. Green & Robert H. Porter, *Noncooperative Collusion Under Imperfect Price Information*, 52 ECONOMETRICA 87 (1984).

[105] Ittoo & Petit, *supra* note 5, at 1. For a skeptical view, see Ulrich Schwalbe, Algorithms, Machine Learning, and Collusion 16 (2018) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3232631 [https://perma.cc/HJ4U-UP3N] ("[C]oordinated behaviour of algorithms is a possible outcome, but it is not as quick and easy or even unavoidable as it is often assumed in the legal discussion of algorithmic collusion.").

[106] Ittoo & Petit, *supra* note 5, at 11–12.

[107] *See supra* notes 89–92 and accompanying text.

[108] *See generally id.* at 13.

[109] *Id.* ("The introduction of Deep RL agents (like Deep Q-Networks) on markets may alleviate some of the obstacles to tacit collusion that we have identified. In particular, Deep RL agents may be quite effective at learning the Q-values of rival oligopolists."); Schwalbe, *supra* note 5, at 3 ("Considering the rapid progress in AI-research [] it cannot be excluded that in the future, algorithms may learn to communicate and to behave in a collusive way.").

3. *Price Discrimination and Coordination*

So far we have assumed that coordinating competitors set similar, although supra-competitive, trade terms for consumers. But in the digital world, another factor comes into play: as more data are gathered about each consumer's preferences, a personalized "digital profile" can be created through the use of algorithms that calculate and update consumers' elasticity of demand in real time.[110] This digital profile can be used by suppliers to increase their profits, by setting the maximal price that each consumer is willing to pay ("personalized pricing").[111] This, in turn, implies that setting one price for all consumers may be welfare-reducing for suppliers, and that more factors must enter into the coordinated equilibrium, making coordination more complicated.

For the purposes of the analysis below, let us assume that personalized pricing can be practiced, even if not to a perfect extent, given factors such as unclear price signals on the part of consumers, unknown demand for new products, and the effects on demand of changing market conditions. How is coordination affected by such opportunities for price discrimination? If no firm has a significant comparative advantage over other competitors, then incentives to engage in coordination may be increased. This is because, without coordination, it will be more difficult to reach a jointly profitable equilibrium.

At the same time, increased information about consumers' real-time preferences also makes it more difficult to coordinate trade terms. The exponential increase in the number of parameters that must be taken into account in calculating personalized prices, as well as in the calculation of a jointly profitable price, introduces "noise" into the system.[112] Furthermore, the ability to coordinate depends, inter alia, on the information about each consumer's preferences held by each supplier.

So, what should be expected? Firms may reach market-division agreements (e.g., Firm A sell to businesses and Firm B sell to individuals), where they all agree not to enter each other's market segment, and each can exploit information regarding consumer preferences in its designated market. Another possibility is that all firms will come to possess similar information, either because consumers' individual preferences are easily calculated, or because all firms refer to a common database and similar data analysis tools. If so, firms can in theory coordinate with respect to the prices charged to each and every

[110] For an example of a digital profile which predicts defaults on loans, see Talia B. Gillis & Jann Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV., at 1 (forthcoming 2019).

[111] *See, e.g.*, Oren Bar-Gill, *Algorithmic Price Discrimination: When Demand Is a Function of Both Preferences and (Mis)Perceptions*, 86 U. CHI. L. REV. (forthcoming 2019); Harrington, *supra* note 6, at 54.

[112] *See, e.g.*, Nicholas Petit, *Antitrust and Artificial Intelligence: A Research Agenda*, 8 J. EUROPEAN COMPETITION L. & PRAC. 361, 361 (2017).

consumer. While such coordination would be almost impossible for humans, it can be facilitated by algorithms under certain market conditions. Alternatively, the difficulties involved in coordination might lead to market equilibriums that, while not fully embracing personalized pricing, would still increase all player's profits under the circumstances.

Observe, however, that the threat of personalized pricing might not be as significant as some claim, for two business-related reasons. First, as Amazon learned the hard way, personalized pricing might create a public backlash.[113] Second, and relatedly, in order to avoid personalized pricing, consumers might prefer to browse anonymously. This, in turn, will limit sellers' ability to engage in targeted advertising. The financial loss from the reduced ability to better identify those potential consumers who might buy a product might well be larger than the loss from not being able to perform personalized pricing. When this is true, personalized pricing will not be practiced.

A related issue involves the use of consumers' digital profiles to individualize products to better meet the preferences of different consumers. This, in turn, may lead to product heterogeneity, which makes coordination harder to sustain. The same observations made above apply here as well. Undoubtedly, a focal point on which to base a coordinated equilibrium may be more difficult to find where differentiated products are offered. Yet algorithms may ease this difficulty—even if not erase it—by engaging in a quicker and more accurate multi-factored analysis.

4. *Algorithms and Harm to Welfare*

Undoubtedly, the effects of algorithms on coordination should be studied further by economists and computer scientists. Yet the potential effects of algorithmic-facilitated coordination are too significant to be ignored until such detailed studies are performed. The analysis presented above, detailing how the characteristics of algorithms operating in the digital economy can, under certain circumstances, facilitate coordination and guide the development of a legal framework aimed at addressing this issue.

By way of summary, I relate briefly to claims raised by some researchers that algorithms do not create significant concerns. Ulrich Schwalbe, Professor of Economics at the University of Hohenheim, argues that "it is doubtful whether algorithms raise barriers to entry,"[114] the fourth condition necessary

[113] *Test of "dynamic pricing" angers Amazon customers*, WASH. POST (Oct. 7, 2000), http://www.citi.columbia.edu/B8210/read10/Amazon%20Dynamic%20Pricing%20Angers%20Customers.pdf [https://perma.cc/VV5H-5EPU]. Nonetheless, the tolerance of consumers to price discrimination may change, as it becomes more prevalent, or once it is connected with personalized (rather than homogenous) products.

[114] Schwalbe, *supra* note 5, at 4.

for coordination. As noted above, I generally agree with this claim. Yet it does not lead to a conclusion that algorithms do not matter. Rather, in markets in which entry barriers are high, algorithms can make coordination easier.

It may be claimed that the fact that, thus far, only a small number of cases involving algorithmic-facilitated cartels have been brought by competition authorities indicates that algorithms have no significant effect. Yet low levels of current enforcement may not reflect market behavior, given that enforcement agencies have only begun to wrap their heads around this new technological challenge, which may require adding computer scientists to their teams. Alternatively, current levels of enforcement may signify that market participants have only recently begun to experiment with the use of algorithms to set prices. Furthermore, it may indicate, as elaborated in the next Part, that legal tools are insufficient to capture some instances of algorithmic-facilitated coordination.[115] Whatever the reason, given that both theory and experimental evidence already point to the potential coordination-facilitating capabilities of algorithms, it is urgent that we prepare for such algorithmic interactions.[116]

A related claim is that none of the cases feature human-less implicit coordination, and that in those cases that were brought, algorithmic technology simply removed the last obstacle to it.[117] While autonomous coordination is probably the most theoretically intriguing scenario, cases in which algorithms tilt the balance towards coordination, because all other market conditions conducive to coordination already exist, should not be treated lightly. Their effects, compared to markets without algorithms, may well be significant. And given the exponential growth in our understanding and applications of machine learning, we cannot afford to wait until algorithms become completely autonomous to check whether our laws are welfare-enhancing.

Some argue that algorithms have difficulties in meeting the need to communicate, which is a fundamental requirement for coordination.[118] While communication is indeed a condition for coordination, as elaborated above, the characteristics of algorithms, and the digital world in which they operate, create communication. Algorithms are "recipes for future action" that increase clarity of how trade terms will be set by them, and how they will react to their competitors' terms.[119] By enabling other algorithms to "read their minds"—either directly or indirectly, even before any action was taken by them—they

[115] *See infra* Part III.

[116] *See* Harrington, *supra* note 6, at 69.

[117] Ittoo & Petit, *supra* note 5, at 2–3.

[118] *Id.* at 3.

[119] *See* Von Neumann, *supra* note 89; *see also* Harrington, *supra* note 6, at 46–47.

limit the need for direct communication or physical meetings. Also, due to the conditions in the digital world, there is lesser need for communication ex ante. Rather, algorithms can coordinate actions in a short sequence of low-value games.[120]

Another claim is that coordination is more difficult to achieve as algorithms become more and more sophisticated.[121] The level of sophistication of an algorithm is determined by those employing it. Furthermore, as the example above indicated, algorithms can be simple. Moreover, sophisticated analysis, which relates to changing market conditions, can strengthen the equilibrium, rather than weaken it.

So how do we ensure that welfare is increased in the data-driven algorithmic economy? What follows is an exploration of two potential tools to limit the negative effects of algorithmic-facilitated coordination: market-based solutions and antitrust law.

III. MARKET-BASED SOLUTIONS

Can the market devise its own solutions to algorithmic coordination? The answer is a partial yes. As shown by Gal and Elkin-Koren, the use of algorithms by consumers can counteract at least some of the effects of algorithmic-facilitated coordination by suppliers. [122] Put differently, it sometimes takes a (consumer) algorithm to beat a (supplier) algorithm.

Algorithmic consumers (digital butlers) are algorithms employed by consumers which make and execute decisions for the consumer by directly communicating with other systems through the Internet.[123] The algorithm automatically identifies a need, searches for an optimal purchase, and executes the transaction on behalf of the consumer. Such algorithms can significantly reduce search and transaction costs, overcome biases, and enable more rational and sophisticated choices.[124] The analysis below assumes that algorithmic consumers are coded to best serve the consumer. This assumption is relaxed later on.

Algorithmic consumers are already part of our digital marketplace. In some industries, such as stock trading, algorithms automatically translate their results into buying decisions;[125] consumers can already purchase a washing machine

[120] *See* Tennenholtz, *supra* note92.

[121] Ittoo & Petit, *supra* note 5, at 2.

[122] *See* Gal & Elkin-Koren, *supra* note 3, at 331.

[123] *See id.* at 313.

[124] *See id.* 313–15.

[125] *See* Shobhit Seth, *Basics of Algorithmic Trading: Concepts and Examples*, INVESTOPEDIA (Mar. 14, 2019), https://www.investopedia.com/articles/active-trading/101014/basics-

that automatically restocks detergent;[126] and a British application monitors prices in the energy market and automatically switches suppliers when it is profitable to do so.[127] Scientists envisage that in the near future algorithmic consumers will become the rule rather than exception for an exponentially increasing number of transactions—realizing a vision of a world where "humans do less thinking when it comes to the small decisions that make up daily life."[128]

Algorithmic consumers have the potential to counteract at least some of the negative welfare effects of algorithmic-facilitated supplier coordination. The Section below explores several such ways, all based on the idea that instead of passively accepting suppliers' decisions, consumers take the reins and actively change market conditions.

Algorithmic consumers can create buyer power if a sufficiently large number of consumers use a specific algorithm, or if several algorithmic consumers coordinate their conduct.[129] This, in turn, may allow consumers to counteract the power of suppliers. The aggregation of consumers can also make transactions larger and less frequent, thereby increasing suppliers' incentives to deviate from the coordinated equilibrium,[130] or to transact "off the digital grid." Such negotiations need not necessarily involve human intervention.

Algorithmic consumers can also be coded to include decisional parameters designed to eliminate, or at least reduce, some market failures.[131] Algorithms are sufficiently flexible to include considerations such as long-run effects on market structures that might harm consumers. For example, an algorithm might be able to recognize coordination among suppliers and refrain from doing business with these suppliers until prices are lowered. Alternatively, to strengthen incentives for new suppliers to enter the market, the algorithm might be coded to always buy some portion of certain goods from at least one

algorithmic-trading-concepts-and-examples.asp [https://perma.cc/5AWM-7MR3]; *Algorithmic Trading*, WIKIPEDIA, https://en.wikipedia.org/wiki/Algorithmic_trading [https://perma.cc/EL8D-64G4].

[126] IBM INST. FOR BUS. VALUE, ADEPT: AN IOT PRACTITIONER PERSPECTIVE, DRAFT COPY FOR ADVANCED REVIEW 13 (2015).

[127] FLIPPER, https://flipper.community/ [https://perma.cc/D6QW-6HZU] (last visited Mar. 20, 2019).

[128] Danny Yadron, *Google Assistant Takes on Amazon and Apple to Be the Ultimate Digital Butler*, GUARDIAN (May 18, 2016), https://www.theguardian.com/technology/2016/may/18/google-home-assistant-amazon-echo-apple-siri [https://perma.cc/7VQC-ADEW].

[129] *See* Gal & Elkin-Koren, *supra* note 3, at 311.

[130] *See id.* at 330.

[131] *Id.*

new source. Of course, including such decisional parameters requires sophisticated modeling and analysis of market conditions, but given ongoing advances in data science, this will become easier.[132] It also requires incentives for collective action, given that refraining from doing business with certain suppliers may be personally costly to individual customers, while disrupting coordination is a public good that benefits all customers since it may eventually lead to lower prices. Such incentives can be created when many consumers are aggregated through an algorithmic consumer.

Finally, algorithmic consumers may reduce the ability of suppliers to engage in personalized pricing.[133] By aggregating the choices of different consumers into one virtual buyer, algorithmic consumers can obscure consumers' personal demand curves (what might be called "anonymization-through-aggregation").[134] More precisely, if consumers are aggregated into sufficiently large consumer groups, suppliers lose the ability to collect data on consumers' individual preferences and to discriminate among them.

In short, algorithmic consumers can potentially improve market dynamics and limit the harmful effects of algorithmic-facilitated supplier coordination without need for legal intervention. Rather, their regulating power resides in the proactive actions of consumers.

This market-based solution is not, however, a panacea. Three main potential limitations can be identified. First, the use of algorithmic consumers may itself infringe on antitrust laws, if they are found to engage in anti-competitive agreements or to abuse their market power.[135] Therefore, it is important to clarify the rules that will be applied to the use of buyer power to counteract supplier power.[136] The second concern is that the market for algorithmic consumers could be dominated by digital butlers (such as Amazon's Alexa) that are not benign, but rather serve the purposes of their suppliers.[137] Indeed, the major digital platform owners are already vigorously competing in the supply of digital assistants.[138] As observed by Ezrachi and

[132] *See, e.g.*, Ittoo & Petit, *supra* note 5; Schwalbe, *supra* note 5.

[133] Gal & Elkin-Koren, *supra* note 3, at 311 ("We do not relate to the welfare effects of price discrimination.").

[134] *Id.*

[135] *Id.* at 345.

[136] On the regulation of buyer power, see generally PETER C. CARSTENSEN, COMPETITION POLICY AND THE CONTROL OF BUYER POWER (2017).

[137] *See* EZRACHI & STUCKE, *supra* note 5, at 191–92.

[138] *See* Mark Prigg, *Apple Unleashes Its AI: 'Super Siri' Will Battle Amazon, Facebook and Google in Smart Assistant Wars*, DAILY MAIL (June 13, 2016), http://www.dailymail.co.uk/sciencetech/article-3639325/Apple-unveil-SuperSiri-Amazon-Google-smart-assistant-wars.html [https://perma.cc/P67F-PDL9].

Stucke, their incentives to do so are straightforward: digital assistants are likely to become consumers' gateway into the digitized world.[139] This, in turn, strengthens the incentives of current platform owners to pursue dominance in the market for algorithmic consumers.[140] Finally, suppliers may take actions to limit the operation of algorithmic consumers.

Other market solutions may also limit the ability of suppliers to engage in algorithmic-facilitated coordination. For example, digital literacy, which ensures that consumers know their options and understand how supplier algorithms work and interoperate, may affect consumer choices.[141] Yet market solutions are, at best, partial. Furthermore, consumers might not be aware that prices are supra-competitive or that their suppliers coordinate their prices. Accordingly, I now turn to legal solutions that can complement or support such market solutions.

IV. LEGAL SOLUTIONS: ALGORITHMIC INTERACTIONS AS AGREEMENTS IN RESTRAINT OF TRADE?

"Smart coordination" by suppliers requires "smart regulation"—setting rules that limit the harms of increased coordination while ensuring that the digital economy's welfare-enhancing effects are not lost.[142] The question is whether antitrust law, which deals with anti-competitive conduct, is fit for the task.[143] This question arises because current legal tools were designed to deal with human facilitation of coordination. New and improved ways to coordinate, as well as the potential scale and scope of the resulting conduct, were not envisioned when antitrust prohibitions were fashioned. It is necessary to determine whether algorithmic interactions that lead to price coordination can and should be caught under existing laws, and if so, to what extent.

Antitrust law currently relies on the exploitation of human limitations in order to increase competition in the market. For example, it prevents market

[139] *See* EZRACHI & STUCKE, *supra* note 5, at 191–92.

[140] *See id.*

[141] Michal S. Gal, *Algorithmic Challenges to Autonomous Choice*, MICH. TELECOMMS. & TECH. L. REV. (forthcoming 2019).

[142] For a similar suggestion, see EZRACHI & STUCKE, *supra* note 5; OECD, *supra* note 5, at 46–47.

[143] For some discussions of this issue, see, e.g., EZRACHI & STUCKE, *supra* note 5; OECD, *supra* note 5; Directorate for Fin. & Enter. Affairs Competition Comm., *Algorithms and Collusion – Summaries of Contributions* (June 2017) (summarizing each country's contributions to the OECD Roundtable on Algorithms and Collusion); Peter Picht & Benedikt Freund, *Competition (Law) in the Era of Algorithms* (Max Planck Inst. for Innovation & Competition Research Paper No. 18-10, May 15, 2018), https://ssrn.com/abstract=3180550 [https://perma.cc/K9SY-4ZQP].

players from discussing anti-competitive agreements and from using the legal system to implement them in order to make it harder to reach and enforce such agreements.[144] But in the algorithmic world, where coordination, detection, and punishment are automated, questions of reaching or enforcing explicit agreements fall in importance. Similarly, the law is based on the assumption that humans' capacity to respond quickly to market changes is limited when numerous or multi-factored decisions must be taken; algorithms are only limited by their computational powers. Furthermore, the current legal treatment of illegal agreements is generally focused on the means of communication used by market players in order to coordinate.[145] When means of communication change, the law might no longer capture conduct which is socially harmful. The challenge is, therefore, to determine to what extent we can rely on existing laws in order to prevent new ways of engaging in socially harmful anti-competitive conduct. More fundamentally, given changes in modes of communication, which may facilitate many more instances of conscious parallelism, we need to explore whether it is still socially beneficial to consider such conduct to be legal. The answers to these questions also serve as a basis for exploring whether new regulatory tools are needed.

The analysis below focuses on how to apply the prohibition of agreements in restraint of trade to algorithms that facilitate coordination. For liability to arise, market participants must be found to have engaged in an agreement which restrains trade, with no offsetting procompetitive effects.[146] The application of additional existing regulatory tools, such as those designed for shared monopolies and merger reviews, is left for future research. Accordingly, the analysis below strives to explore and provide preliminary answers to two interconnected questions that stand at the basis of designing a welfare-enhancing policy toward the use of coordination-facilitating algorithms:

1. Do algorithms that facilitate coordination fulfill the requirement for "an agreement" as defined in antitrust laws, and, if so, under what conditions?
2. If the answer to the first question is positive, what exactly do we wish to prohibit, and can we spell this out clearly for market participants?

The answer to the first question is quite often positive. The real challenge lies in the second question, which focuses on whether and under what

[144] *See, e.g.*, Harrington, *supra* note 6, at 46–47.

[145] *See, e.g.*, Page, *supra* note 17, at 599–601; Kaplow, *supra* note 20; Harrington, *supra* note 6, at 46–47.

[146] *See* Sherman Anti-Trust Act, 15 U.S.C. § 1 (2018).

conditions algorithms should be treated as engaging in "restraint of trade." The answers to these questions also depend on our ability to set rules that can also be justified based on decision-theory considerations,[147] ensuring that the actual costs of enforcement do not outweigh its benefits given institutional limitations.

One last general note is in order. It is important to separate two questions that arise: whether an illegal agreement has been reached, and who is legally liable for it. This Article focuses on the former.

A. Coordination-Facilitating Algorithms as "Agreements"

1. *General: Agreement, Plus Factors and Facilitating Practices*

For liability to arise from coordinated conduct, an "agreement" must be found to exist.[148] But what is an agreement? Despite the importance of this concept and the numerous cases and commentary which have strived to define it, the term's meaning remains vague and its boundaries are contested.[149] Yet some principles are largely agreed upon. As the Supreme Court noted in *Bell Atlantic Corp. v. Twombly*, an agreement must involve either express or tacit (i.e., implicit) formulation.[150] Independent conduct, in which competitors act in parallel without regard to one another's actions, does not constitute agreement, nor does mere interdependent conduct (conscious parallelism), in which firms take into account how other firms are expected to react.[151]

Despite wide agreement on these principles, some prominent scholars suggest that the term "agreement" is sufficiently broad to capture conscious parallelism. This argument was famously raised (though recently repudiated) by Richard Posner,[152] who argued that conscious parallelism involves the making and acceptance of an offer through conduct, and therefore, literally and materially fulfills the conditions for an agreement. This view, dormant for many years, was recently endorsed by Harvard University Law Professor Louis

[147] On decision-theory in antitrust, see, e.g., C. Frederick Beckner III & Steven C. Salop, *Decision Theory and Antitrust Rules*, 67 ANTITRUST L.J. 41 (1999).

[148] The word "agreement" is used broadly to include alternative wordings (e.g., arrangement).

[149] Contrast, for example, KAPLOW, *supra* note 20; Page, *supra* note 17; Kovacic, *supra* note 94.

[150] Bell Atl. Corp. v. Twombly, 550 U.S. 544 (2007); *see* William H. Page, *Tacit Agreement Under Section 1 of the Sherman Act*, 81 ANTITRUST L. J. 201, 209–10. The use of the term "tacit agreement" is confusing, since it is sometimes used to indicate conscious parallelism. I assume that the Court intended to differentiate between these terms.

[151] *Id.* at 601–02.

[152] RICHARD A. POSNER, ANTITRUST LAW: AN ECONOMIC PERSPECTIVE 146 (1976). More recently, Posner repudiated his view. Richard A. Posner, *Review of Kaplow, Competition Policy and Price Fixing*, 79 ANTITRUST L.J. 761, 766 (2014).

Kaplow.[153] Analyzing economic models as well as United States' case law, Kaplow makes a strong and convincing case that the distinction between express collusion and conscious parallelism is blurry, and the definition of "agreement" can include both.[154] Furthermore, he shows that some Supreme Court precedents are sufficiently wide as to be interpreted to include conscious parallelism.[155] He also argues that the distinction between the two does not serve social welfare. The main problem with this view lies in the practical limitations of prohibiting conscious parallelism. Indeed, the problem of fashioning a clear prohibition and an applicable remedy has been one of the main reasons for treating conscious parallelism as legal.[156] Kaplow addresses this problem by suggesting that the prohibition be structured to incentivize market participants to act as if in a one-shot game without fines, which would lead to competitive prices. He also argues that if the remedy is sufficiently strong, market players will have sufficiently strong motivations not to engage in the prohibited conduct.[157] However, practical questions still remain: how to clarify what conduct is prohibited, and whether courts can readily apply such a prohibition in practice. Posner recently acknowledged these problems, citing them as a reason for repudiating his earlier views.[158] For the purpose of this Article, I assume that conscious parallelism is not currently captured by the law.

The focus thus shifts to the definition of tacit agreements, which come under the law. This concept is not clearly defined.[159] Its name indicates that an agreement is implied or indicated, but not explicitly expressed.[160] While clearly some form of meeting of minds is necessary, neither the law nor Supreme Court precedents clearly clarify what constitutes an illegal meeting of minds that could be differentiated from the meeting of minds that stands at the basis of conscious parallelism. In both cases, the parties take into account the

[153] KAPLOW, *supra* note 20, at 77–82.

[154] *Id.*

[155] *Id.* With regard to the relevance of *Twombly*, it is argued that as the Court did not carefully articulate any concept of agreement, reconcile its statements with prior conflicting statements, or discuss its reasons for its interpretation, its relevance should be limited. *Id.* at 88–92.

[156] *See, e.g*, Donald F. Turner, *The Definition of Agreement Under the Sherman Act: Conscious Parallelism and Refusals to Deal*, 75 HARV. L. REV. 655, 657–84 (1962).

[157] *See* KAPLOW, *supra* note 20.

[158] Posner, *supra* note 152, at 766.

[159] Furthermore, it confused matters further by stating that "allegations of parallel conduct . . . must be placed in a context that raises a suggestion of a preceding agreement." Bell Atlantic Corp. v. Twombly, 550 U.S. 544, 557 (2007). This implies that the parties have already formed an express agreement, which they then implement.

[160] KAPLOW, *supra* note 20, at 36.

expected reactions of their competitors; in both, some flow of information is necessary; in both, there must be intent to engage in coordinated conduct.

Most commentators and courts suggest a definition that focuses on communication between competitors which signal intent to act in a coordinated way, and their reliance on each other to follow suit.[161] The mode of communication, as well as the types of information communicated, play a decisive role under such definitions. Building on lower court precedents, University of Florida Law Professor William Page suggests that tacit agreement be defined to include two-staged situations in which competitors "clarify their expectations about one another's intentions by communication, then act consistently with the communications."[162] No exchange of express assurances to act uniformly is required. [163] An additional requirement is that communication take place by means that lack efficiency justifications.[164] This condition ensures that the communication would not have taken place regardless of its coordinating effects, and it reduces the risk that deterring the communication will harm social welfare.

To assist in separating conscious parallelism from tacit agreement, lower courts have endorsed the concept of "plus factors"—i.e., circumstantial facts or factors that go beyond mere conscious parallelism, from which an agreement can be indirectly inferred.[165] Plus factors can be negative or positive. Negative plus factors are the fruits of economic reverse-engineering: absent an agreement, it is improbable that parallel conduct would have arisen under the given market conditions.[166] Since parallel conduct took place, it can thus be inferred that an agreement was reached between market participants. Similar bids for made-to-order products exemplify this category: they could not have occurred absent prior agreement among the bidders. Interestingly, algorithms make proving the existence of negative plus factors more difficult. This is because their characteristics make it easier to reach parallel conduct without an agreement. This, in turn, increases what Kaplow calls the "paradox of proof":

[161] *See* Kaplow, *supra* note 20; Harrington, *supra* note 6, at 25–46. Interestingly, Professor Harington suggests that overt communication is not a necessary part of a collusive scheme, which he defines as when firms use strategies that embody a reward-punishment scheme which rewards a firm for abiding by the supra-competitive outcome and punishes it for departing from it. Yet the requirement for communication reduces false positives and serves as an informative signal for the presence of collusion. *See id.*

[162] Page, *supra* note 17, at 608.

[163] *Id.*

[164] *Id.*

[165] *See, e.g.*, OECD, *supra* note 5, at 20; Kovacic et al., *supra* note 94.

[166] *See* AREEDA & HOVENKAMP, *supra* note 13, at 181–82; Harrington, *supra* note 6, at 27 (unnatural parallelism).

the more conducive are existing natural market conditions to coordination, which makes price elevation and the resultant harm to social welfare more likely, the less the need for specified means of communication such as those currently required to prove an agreement, and the lower the chance that an agreement will be proven and the conduct condemned.[167]

Positive plus factors constitute avoidable acts that indirectly prove a shared commitment to a common cause.[168] Yet the scope of application of this requirement is unclear and sometimes misleading.[169] Some examples of plus factors used by courts can as readily indicate conscious parallelism, and therefore add to the confusion. For example, "acts against one's self-interest," which make sense only if we read them to include acts against one's short-term interests, also characterize conscious parallelism: a competitor does not lower its price below the jointly profitable level, even though it can profit in the short run, because it acknowledges that such an action might trigger retaliation by its competitors, which would lower its profits in the long run.[170]

Other examples are less problematic. These include, for example, meetings of competitors without other justifications, and private disclosure of future price changes.[171] Importantly for our discussion below, while it is settled law that "the *form* of [communication] should not be determinative of its legality[,]" [172] the requirement that the communication lack efficiency justifications has made many courts reluctant to find an agreement when the communication is public and relates to current or future trade terms.[173] Public price announcements have generally been treated as creating transparency for consumers as well as shareholders.[174] Some courts put heavy emphasis on pre-action explicit communication of promises to act in a certain way and threats to punish deviations.[175]

[167] KAPLOW, *supra* note 20, at 124–73.

[168] While courts vary with regard to the scope of the concept, some core examples of plus factors are widely accepted. Compare, for example, Kovacic et al., *supra* note 94; RICHARD A. POSNER, ANTITRUST LAW 55–93 (2nd ed. 2001); KAPLOW, *supra* note 20.

[169] *See, e.g.*, KAPLOW, *supra* note 20, at 111–14.

[170] *See id.* at 111.

[171] *See* Page, *supra* note 17, at 221.

[172] *In re* Coordinated Pretrial Proceedings in Petroleum Prods. Antitrust Litig., 906 F.2d 432, 447 (9th Cir. 1990).

[173] Dennis W. Carlton, Robeert H. Gertner & Andrew M. Rosenfeld, *Communication Among Competitors: Game Theory and Antitrust*, 5 GEO. MASON L. REV. 423, 428–29 (1997) (communications are most likely to be anticompetitive if they are private rather than public, if they relate to current and future prices rather than historical prices, and are repeated rather than isolated).

[174] *Id.* at 432.

[175] *See* discussion in Page, *supra* note 17; Kovacic et al., *supra* note 94.

The related concept of facilitating practices is also relevant to our discussion. Facilitating practices are positive, avoidable actions that allow competitors to more easily and effectively achieve coordination by overcoming impediments to coordination, in a way that goes beyond mere interdependence.[176] In doing so, they increase competitors' incentives to cooperate, despite their divergent interests.[177]

When firms expressly agree to adopt a facilitating practice—for example, agreeing to post their prices in advance—that agreement, by itself, may constitute an agreement in restraint of trade.[178] More relevant to our discussion are instances under which facilitating practices themselves are prohibited. Toward this end, two main (and partially overlapping) legal routes are possible.[179] The first treats the adoption of facilitating practices, by itself, as a basis for liability. This route was first suggested by the late Harvard University Professor Donald Turner but was never adopted.[180] As elaborated below, given the shortcomings of existing law in addressing algorithmic-facilitated coordination, the time may be ripe to rethink this position. The second route, which is currently applied, treats the adoption of facilitating practices as a sub-category of plus factors: under certain circumstances they serve as indirect indications of an "agreement."[181] Both legal routes recognize that a facilitating practice can also create procompetitive effects, such as providing consumers and potential entrants with more accurate information necessary for their decisions.[182] Therefore, both also include tools designed to ensure that procompetitive justifications are included in the analysis. Yet they are conceptually different. The former prohibits the conduct itself, given its potential anticompetitive tendencies. The latter is evidentiary: the use of facilitating practices serves as an indirect circumstantial indication of an agreement between parties operating in the market.

The logic behind the existing rule can be explained as follows. Facilitating practices are avoidable actions which change market conditions in a way that

[176] *See, e.g.*, Steven C. Salop, *Practices that (Credibly) Facilitate Oligopoly Coordination*, *in* NEW DEVELOPMENTS IN THE ANALYSIS OF MARKET STRUCTURE 265, 271 (Joseph Stiglitz & G. Frank Mathewson eds., 1985); Charles A. Holt & David T. Scheffman, *Facilitating Practices: The Effects of Advance Notice and Best-Price Policies*, 18 RAND J. ECON. 187 (1987); Ian Ayres, *How Cartels Punish: A Structural Theory of Self-Enforcing Collusion*, 87 COLUM. L. REV. 295 (1987); Page, *supra* note 40; KAPLOW, *supra* note 20, at 276–85.

[177] Salop, *supra* note 176, at 434–35.

[178] *Id.* 425–26.

[179] KAPLOW, *supra* note 20, at 276.

[180] Turner, *supra* note 156, at 666–67.

[181] Page, *supra* note 20, at 415–16.

[182] *Id.*

makes it easier to coordinate. In the absence of procompetitive justifications for their adoption, firms would not have engaged in such conduct unless they served as an indirect communication device to signal to each other their intent to engage in coordinated conduct and their reliance on their competitors' acceptance of such practices. Accordingly, the facilitating practice provides indirect proof of an "agreement."

Many facilitating practices exist, with varying degrees of success in promoting coordinated conduct.[183] Steven Salop identifies two distinct types: information exchange and incentive management.[184] Information exchange devices facilitate coordination by reducing uncertainty about competitors' actions and intentions.[185] For example, sharing information on actual sales and costs may enable competitors to determine whether a price reduction represents an instance of defection. Incentive management devices alter the structure of firms' pay-off matrices, thereby affecting their incentive to offer price discounts.[186] Meeting competition clauses illustrate this effect. Under meeting competition clauses, a firm announces that its price will not be higher than the lowest price posted by another firm.[187] Such clauses automatically incorporate the aggressive response to price-cutting—i.e., immediate price matching—needed to support coordination. Consumers are used to police the agreement, because the risk of missing out on the lowest price creates incentives for them to assume the costs of monitoring suppliers' conduct. These clauses may not be in consumers' interest if their collective acceptance stabilizes suppliers' joint profit outcomes and makes discounting less desirable.[188]

In today's digital world, there is less need for some information-exchange facilitating practices. Real-time data collection and rapid analysis make information exchange agreements redundant if relevant data can be easily collected through independent means. Still, other forms of information exchange may facilitate coordination, such as those pertaining to the kinds of datasets used by an algorithm, competitors' output and cost data, or the decisional parameters included in the algorithm.[189] With respect to incentive management devices, some may be even more potent in the digital world. Take, for example, meeting competition clauses, in which the online retailer promises consumers it will meet any lower price found on the Internet. If

[183] Salop, *supra* note 176.

[184] *Id.*

[185] *Id.* at 272.

[186] *Id.*

[187] *Id.* at 280.

[188] *Id.* at 273.

[189] Reverse-engineering or backtracking logic can sometimes be used to determine such data without information exchange.

lower prices are immediately matched, competitors have no incentive to offer a discount.

2. *Applications of the Concepts to Algorithms*

Let us now relate the above concepts to algorithmic interactions. As will be shown, some concepts are as relevant as ever, while others are challenged by the digital world. The difficulty arises from the discord between existing conceptions and assumptions—shaped to apply to human interactions—and the way in which the digital world operates.

Some types of coordination between algorithms easily fall within the definition of agreement. A simple scenario involves the use of algorithms to implement, monitor, police, or strengthen a prior, explicit agreement among suppliers. In such situations, a clear agreement exists between the users of the algorithms, and the algorithms simply serve as the tools for their execution.[190] The case brought in 2015 by the U.S. Department of Justice against David Topkins for coordinating with other sellers the prices of posters sold online, illustrates such agreements. Topkins and his co-conspirators designed and shared dynamic pricing algorithms, which were programmed to act in conformity with their agreement.[191] The algorithms played a secondary role, based on an existing agreement between the sellers.[192] Such use of algorithms is not much different from a previously agreed upon price formula, even if the algorithm determines the final price based on such a formula, and takes into account data on market conditions inputted into it at any given time. FTC Commissioner Maureen Ohlhausen suggested a simple test that captures many of these easy cases: if the word "algorithm" can be replaced by the phrase "a guy named Bob," then algorithms can be dealt with in the same way as traditional agreements.[193]

The more difficult cases arise when algorithms are designed independently by market players to include decisional parameters that react to other players' decisions in a way which strengthens or maintains a joint coordinated outcome.[194] For example, a programmer might base the algorithm's decisional parameters on his predictions of the best responses to other players' conduct (an "expected coordination algorithm"). The algorithms explored in detail in the previous Section illustrate this case: They are designed and adopted

[190] For four main scenarios, see EZRACHI & STUCKE, *supra* note 5.

[191] Press Release, U.S. Dep't of Justice, Office of Pub. Affairs, Former E-Commerce Executive Charged with Price Fixing in the Antitrust Division's First Online Marketplace Prosecution (Apr. 6, 2015) [hereinafter Topkins Press Release].

[192] *See id.*

[193] *See* Ohlhausen, *supra* note 12.

[194] *See* EZRACHI & STUCKE, *supra* note 5.

independently, without prior meetings or commitments, but each player independently codes his algorithm so that it takes into account other players' probable reactions, as well as their joint incentive to cooperate.[195] Even more difficult questions arise when algorithms are not deliberately designed in a way that facilitates coordination, yet they autonomously reach the same result. In these cases, the algorithm is given a general goal, such as "maximize profits," and it determines the decisional parameters it will use based on machine learning ("learned coordination").[196] While the question of who is legally liable for coordination may differ between the two scenarios, the two raise the same basic question of whether they reflect the existence of an agreement in the antitrust sense. I therefore explore whether such conduct constitutes (legal) conscious parallelism or (illegal) tacit agreement.

Let us start with the following suggestion: Conscious parallelism that results from algorithms simply mimicking human conduct, making the same decisions and taking the same actions as humans engaged in lawful conscious parallelism, without further facilitating coordination, should not constitute an agreement. [197] Any other rule would unjustifiably differentiate between algorithms and humans. The following example illustrates this point: assume a market in which longstanding conscious parallelism exists. Each of the firms operating in the market adopts an algorithm based on the benchmark for pricing that the firm has been using for years. Does the fact that market players are now using algorithms to achieve an identical result change the legal status of their conduct? If each supplier unilaterally and independently decides to adopt such an algorithm, and the algorithm does not significantly change their ability to reach and maintain the existing jointly profitable equilibrium, then it should not be regarded differently from the original method for decision-making, which was deemed to be legal.[198]

A tougher question arises when the algorithm uses similar decisional parameters and makes similar decisions to those made by humans under a given set of conditions, but in a much more efficient manner, thereby essentially facilitating coordination. Take, for example, the task of detecting price deviations and changing one's price accordingly. Algorithms can more easily perform this task than humans. Should their higher level of efficiency in performing this coordination-strengthening act change its legality? Put

[195] *See* EZRACHI & STUCKE, *supra* note 5.

[196] *Id.*

[197] *See id.*; Harrington, *supra* note 6, at 32.

[198] Harrington, *supra* note 6, at 45–46. Harrington suggests that some pricing algorithms that "condition play on a competitor's past prices" should be prohibited per se under section 5 of the Federal Trade Commission Act. Price matching algorithms would most likely fall under this prohibition. *See id.*

differently, can the employment of the algorithm be treated as a facilitating practice under existing law? The question arises because while the *pattern* of conduct is similar to what would otherwise be considered lawful, the *method* and *effect* of the conduct may differ significantly. As elaborated above, the use of algorithms may strengthen not only the ability, but also the incentives, to coordinate. Moreover, if the algorithm is transparent, it serves, by its nature, as a clear declaration about how the firm is going to react to market conditions, thereby changing the dynamics of the interaction.

Below I analyze the application to algorithmic interactions of some of the requirements, assumptions, and concepts on which antitrust law is based. As will be shown, while the use of algorithms is not prohibited, certain ways of using algorithms or other practices that in combination with algorithms facilitate coordination, may be considered illegal.

a) Application of Basic Concepts

Let us first examine the application to algorithmic interactions of fundamental concepts relating to agreements. This Section argues that the existing taxonomy is generally sufficiently broad as to capture such interactions. Note that at this stage I only explore whether an agreement was formed, not whether it is legal.

Engaging in an agreement requires the intent to do so.[199] Algorithms cannot have a mental state of "intent," or any mental state, for that matter.[200] Yet it might be claimed that algorithms intend to reach a certain goal by using a certain strategy, including reaching a coordinated equilibrium with other algorithms. If we do not wish to go so far, the intent of the programmer to create coordination through the use of algorithms, and the intent of the user to employ such an algorithm, can fulfill this requirement. This is because the algorithm serves as a tool for carrying out the intent of its programmer or user. Some cases are simple, such as the expected coordination scenario, in which the decision to include coordination-facilitating elements in the algorithm is a conscious one.[201] But this may not always be the case. Users may simply not be interested in the parameters which drive the algorithm's decisions. More interestingly, in the learned coordination scenario, the programmer might not be aware of such parameters if the algorithm is based on machine learning.[202] That is, instead of being specifically coded to react in a certain way, an

[199] Courts often focus their analysis on the expressions made by one competitor to another, rather than on intentions. An expression of a willingness to enter into an agreement, even if the competitor had not intent of doing so, suffices. Algorithms can fulfill this requirement. In the European context, see Blockx, *supra* note 86.

[200] Ezrachi & Stucke, *supra* note 47.

[201] *See* Topkins Press Release, *supra* note 186.

[202] *See* EZRACHI & STUCKE, *supra* note 5.

algorithm may be designed such that it independently determines the means to reach a given target through reinforced self-learning. Should the algorithm adopt a strategy that leads to conscious parallelism, coordination will not be the fruit of explicit human design but, rather, the outcome of evolution, self-learning, and independent machine execution.

Can we still find the resulting coordination to be the fruit of a conscious, avoidable act? To our mind, learning algorithms should generally not be treated differently from expert algorithms, which are specifically coded to react in certain ways. While this question deserves an extended analysis, five points are worth making. First, the algorithm's goals are set by its programmer.[203] Indeed, algorithms designed to serve the goals of a particular user act as software agents. These agents may navigate in a computerized network, while transmitting messages among themselves, and interacting with other agents, which might be controlled by other users. Second, algorithms learn from case studies supplied by the programmer and may be reinforced by the programmer's inputs.[204] Third, the programmer can place some limitations on the methods used by the algorithm to make his decisions. At the very least, so long as the algorithm's programmer can code it to *not* act in a certain manner, and incorporate safeguards that limit the scope of its reactions to market conditions (compliance by design), then any programmer's failure to do so should be taken into consideration. This can be likened to limitations placed on autonomous algorithms: self-driving cars should not be able to follow any and all possible decision paths to their logical conclusions simply because their algorithms are autonomous. Furthermore, the treatment of algorithms as a "black box" whose secrets are concealed even to the programmer is fallacious. As Avigdor Gal, Professor of Data Science at the Technion, argues, causal relations between the features (data points) used by an algorithm to reach its decision can be relatively easily observed by the programmer.[205] The programmer can thus be aware of such correlations, at least under certain circumstances. Learning algorithms can thus also be treated as conscious, avoidable acts. Note that this does not imply that such algorithms would necessarily create liability. This question is dealt with in the next Section.

[203] Simonetta Vezzoso, Competition by Design (June 15, 2017) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2986440 [https://perma.cc/WR9H-5RYG]; Emilio Calvano et al., *Algorithmic Pricing: What Implications for Competition Policy?* (July 7, 2018) (unpublished manuscript), https://ssrn.com/abstract=3209781 [https://perma.cc/3LUC-VGEV].

[204] *See* P. Anitha, G. Krithka & Mani Deepak Choudhry, *Machine Learning Techniques for Learning Features of Any Kind of Data: A Case Study*, 3 INT'L J. ADVANCED RES. COMPUTER ENGINEERING & TECH. 4324, 4325.

[205] *See, e.g.*, Gal, *supra* note 53.

Agreement requires a "meeting of minds."[206] Once again, in the expected coordination scenario, the presence or absence of a meeting of minds among the algorithms' programmers or users should determine the fulfillment of this requirement. The learned coordination scenario raises more difficult issues. An algorithm, operated by a computer, does not have a "mind" in the literal sense. Yet it makes decisions based on given inputs, including the expected and actual reactions of others. Moreover, as the studies surveyed above prove, algorithms can autonomously reach coordination which serves their goals.[207] Furthermore, Kaplow suggests that the term meeting of minds "readily covers . . . the standard scenario in which firms in an oligopoly are able to coordinate their prices by understanding each other's thought processes, which forms the basis for predicting their reactions to different prices that each firm may charge."[208] Should this definition be accepted, then it may include cases where algorithmic interactions lead to the conclusion that coordination is their best strategy, given the expected and actual reaction curves of competitors. Finally, the case law suggests that the mere exchange of commercially sensitive information to another party, which influences the action of the recipient, suffices.[209] Algorithms perform this function.

Can an algorithm communicate a conscious commitment to a common theme? Definitely yes. As elaborated above, a transparent algorithm can serve as a recipe for future action, including the price to be paid for deviations, which act as explicit threats of punishment.[210] Employing the algorithm in practice translates such a commitment into actions. While algorithms generally do not sign agreements, wink to each other, or nod their consent, they communicate through the decisional parameters coded into them. Other firms can then rely on such communications in order to shape their own actions.

In the non-algorithmic world, courts often look for evidence tending to show that the defendants "got together and exchanged assurances of common action."[211] Such physical meetings are, obviously, irrelevant to algorithmic interactions. Yet algorithms "get together" in cyberspace. They make use of

[206] *See, e.g.*, discussion in Kaplow, *supra* note 20. For a relatively similar issue in the European context, see, e.g., Andreas Heinemann & Aleksandra Gebicka, *Can Computers Form Cartels? About the Need for European Institutions to Revise the Concertation Doctrine in the Information Age*, 7 J. EUROPEAN COMPETITION L. & PRAC. 431 (2016).

[207] *See* discussion on Page, *supra* note 17.

[208] KAPLOW, *supra* note 20, at 34; *see* Jonathan B. Baker, *Two Sherman Act Section 1 Dilemmas: Parallel Pricing, the Oligopoly Problem, and Contemporary Economic Theory*, 38 ANTITRUST BULL. 143, 178 (1993).

[209] AREEDA & HOVENKAMP, *supra* note 13.

[210] For the role of explicit threats of punishment, see Page, *supra* note 17.

[211] Page, *supra* note 17, at 603.

conditions in the digital world that enable them to observe and react to each other, and that make signaling, information transfer, and exchange of assurances easier.

Should the communication between parties be verbal? Some courts and scholars give weight to verbal communications in their definitions of agreement.[212] Yet it is generally agreed that intentional use of a well-understood nonverbal signal can express assent.[213] Conceptually, the requirement of communication is sufficiently wide as to include all forms of message delivery. Furthermore, mandating a certain kind of communication excludes cases in which competitors reach the same anticompetitive outcome through other means, which could be even more efficient. Accordingly, if exposing an algorithm's decisional parameters sends a signal to competitors, then this should be regarded as communication for legal purposes.

b) Algorithms as Plus Factors

Can the use of algorithms be treated as plus factors—which indirectly prove the existence of an agreement—once parallel conduct is proven to exist? For the answer to be affirmative, their use must constitute an intended and avoidable act that facilitates coordination by creating conscious commitments to a common scheme, which is not justified on procompetitive grounds.[214] Let us apply these conditions to algorithms.

As elaborated in Part II, the design and use of an algorithm is, in itself, an avoidable and intentional act.[215] Such algorithms can facilitate, maintain, or strengthen coordination by limiting incentives to compete beyond those that exist naturally.[216]

Several points are worth emphasizing with regard to the causal connection between the use of algorithms and coordination. First, not all algorithms facilitate coordination. Some may perform functions that do not affect the incentives or ability of firms to coordinate. Obviously, such algorithms should not be considered plus factors. Second, in determining the effects of an algorithm, it is important to separate any facilitating effects of using a given algorithm from facilitating effects that arise from the conditions of the digital world—e.g., increased connectivity. The latter should be taken as a given. Third, the use of algorithms is often combined with other practices that facilitate coordination. For example, a firm might design its website to continually display the price calculated by the algorithm. Or a firm may take

[212] *Id.* at 614–16.

[213] *Id.* at 605–06 and resources cited there.

[214] *See supra* Section IV.A.1

[215] *See supra* Part II.

[216] *See supra* Part II.

measures designed to make the algorithm harder to change, thereby strengthening the degree to which competitors can rely on the algorithm's decisional process. All facilitating practices should be analyzed together. Fourth, it is useful to differentiate between algorithms that facilitate coordination among competitors, and those that might facilitate coordination among other market players. The algorithms used in the online posters case mentioned above illustrate the first case, while price comparison algorithms fall into the second category.[217] These two categories differ in both their economic functions and legal implications. While use of the former may be considered to constitute an agreement, the latter usually cannot.

Another question that arises is whether the adoption of facilitating practices must be uniform. The answer to this question should be negative. Assume, for example, that the algorithms do not employ similar decision trees, but the combination of their decisions nonetheless facilitates coordination. This may be the case when one competitor's algorithm sets a price at the jointly profitable level, and the others set prices based on that algorithm's price (a follower-leader scenario, like the algorithm presented above). In such a situation, requiring adoption of a similar algorithm by all competitors would make it easy to circumvent the requirement of "agreement." Therefore, there is no need for algorithms to be uniform, or for all competitors to employ algorithms, so long as each engages in conscious, avoidable acts that facilitate coordination.

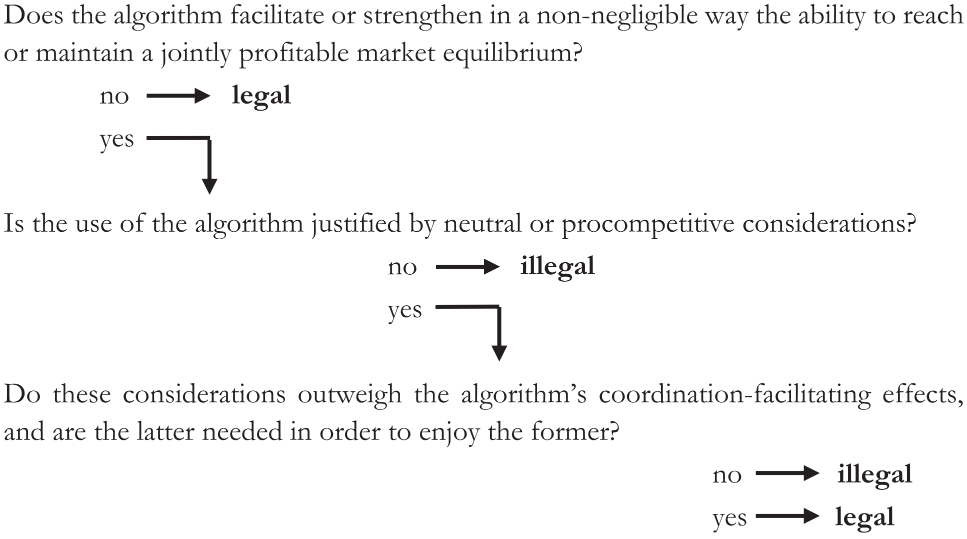
The adoption of certain algorithms, followed by expected accommodating conduct by competitors, can therefore facilitate coordination and imply the existence of an implicit agreement. The problem with treating the adoption of algorithms as plus factors is, however, twofold. First, algorithms perform many functions in the digital environment, and bring about many benefits. Accordingly, if we cast the net too widely, we risk creating a chilling effect on welfare-enhancing conduct. While rules should not allow programmers and users to hide behind algorithms, they should also ensure that what we gain in limiting facilitating practices is greater than what we lose in limiting the range of allowable design choices. This does not imply that we should adopt a "hands off" approach to all algorithms, but rather, we must tread carefully. We should therefore ensure that our laws are based on an understanding of the role of algorithms in the marketplace, including their comparative advantages over human decision-making. In this respect, it makes sense to start with the easy cases in which harm to competition and welfare is more evident.

[217] *See* Topkins Press Release, *supra* note 191; COMPETITION & MARKETS AUTHORITY, DIGITAL COMPARISON TOOLS MARKET STUDY (2017) (UK).

The second problem is the content of the prohibition: what exactly do we wish to prohibit, and can we spell this out clearly for market participants? Can we meaningfully instruct firms how to operate legally? To use Phillip Areeda's suggested rule of thumb: can we indicate, in less than twenty words, what kind of conduct firms are prohibited from engaging in?[218]

In light of the above, the algorithm's ability to facilitate coordination should be balanced against its pro-competitive effects. Algorithms should be subject to the following rule of reason analysis:

Diagram 1: Algorithms as facilitating practices



Does the algorithm facilitate or strengthen in a non-negligible way the ability to reach or maintain a jointly profitable market equilibrium?

no → **legal**

yes ↓

Is the use of the algorithm justified by neutral or procompetitive considerations?

no → **illegal**

yes ↓

Do these considerations outweigh the algorithm's coordination-facilitating effects, and are the latter needed in order to enjoy the former?

no → **illegal**

yes → **legal**

Observe that it should not be necessary for an algorithm to have no potential procompetitive effects—only that the balance should not be tilted toward their anticompetitive outcomes. Otherwise, we might not capture any algorithms under our laws, given that they often create efficiencies.[219] Furthermore, as Kaplow argues, in determining whether a possibly ambiguous practice should be viewed positively or negatively, it is necessary to consider the real effects on the market.[220] If, for example, transparency makes it easier for sellers to identify cheaters and deter defection, then buyers will simply gain better information about high supra-competitive offers.[221] At the same time, it is important to also give weight to wide institutional considerations in order to

[218] Thanks to Bill Kovacic for suggesting the use of this test.

[219] *See* Mehra, *supra* note 5.

[220] KAPLOW, *supra* note 20, at 279.

[221] *Id.* at 279.

ensure that we do not chill efficiency and innovation. This implies that considerations such as creating ex ante certainty should also be weighed.

Importantly, algorithms should not necessarily be treated as indivisible units. Indeed, the facilitating device may form only part of the algorithm. It is often the case that an algorithm performs many functions, such as gathering the data, analyzing it, and determining what trade terms to set based on the data.[222] Many of these functions can be welfare-enhancing, reducing costs or increasing the quality of production or marketing functions, and therefore should be allowed.[223] At the same time, some functions may be used to facilitate coordination. It is thus essential to separate the different functions and determine whether the benefits of the former are dependent on the harms of the latter. Otherwise we risk throwing out the baby with the bathwater. This suggestion also serves as a partial answer to those who are concerned that regulating algorithms would limit the benefits they bring about.

This leads to a third suggestion: because our understanding of how algorithms interact in the digital world is still rudimentary, the rules regulating algorithms should be developed in widening circles, in keeping with our understanding of their potential effects on the market and the potential chilling effects of overbroad prohibition. Accordingly, as a first step, competition authorities should strive to identify the relatively straightforward cases in which the legal requirements can be easily applied and a relatively clear rule can be created.

Below I suggest five cases which raise red flags and therefore are good candidates for a repository of cases characterized by prima facie justification for further examining their legality. All cases share three traits: (1) they may facilitate coordinated conduct; (2) they are potentially avoidable by the algorithm's programmers or users; and (3) they are unlikely to be necessary in order to achieve procompetitive results. Such practices may thus amount to "coordination by design." The cases are as follows:

1. Suppliers consciously use **similar algorithms even when better algorithms are available to them**. The algorithms need not be identical, but their operative part—which calculates the trade conditions—should generate relatively similar outcomes.

Observe that the use of similar algorithms, by itself, is insufficient to lead to a coordinated outcome. This can be illustrated by a simple example: assume that all algorithms base the price on their firm's production costs. If

[222] *See* Org. for Econ. Co-operation & Dev., *Algorithms and Collusion – Background Note by the Secretariat*, OECD Doc. DAF/COMP(2017)4 9–16 (June 6, 2017).

[223] *See id.*

production costs differ among competitors, the algorithms will not lead to a jointly profitable price.

2. Firms make conscious use of **similar data** on relevant market conditions **even when better data sources exist.** Data is an essential input in the decision-making process, which affects the decision. Using similar data is especially important when prices are based on consumers' digital profiles. Note that the data sources themselves need not be identical so long as the information gleaned from them is relatively similar.

3. Programmers or users of learning algorithms give them **similar case studies** from which to learn **despite those not being the best-case studies readily available**. Learning algorithms change their decision trees based on learning from past experience. If fed similar cases, the algorithms may learn similar things and make decisions accordingly.

4. Users take actions that make it **easier for their competitors to observe their algorithms and/or their databases**, and their competitors take actions to observe them. The algorithm can signal to other market players how its user is likely to react to market conditions, thereby communicating intent and possibly a credible commitment.[224] The easiest case arises, of course, when the algorithm is revealed only to one's competitors (either by allowing them to digitally access it or by sending it to them privately). For example, the algorithm might encrypt its information so that only competitors can read it. In such cases it is clear that the algorithm's transparency does not serve consumers, and is artificial rather than an inherent part of digital markets. But even when the algorithm or database is revealed to all, such an action might still amount to a plus factor or a facilitating practice, depending on the circumstances. Those include, inter alia, the following: (1) does such transparency benefit consumers in any significant way; (2) do consumers have the means and incentives to understand the operation of the algorithm; and (3) does the competitor otherwise have incentives to keep the content of the algorithm or the database a trade secret. This category fits well with the current prohibition against the exchange of competitively sensitive information among competitors in an effort to stabilize or control industry pricing.[225]

[224] *See* Harrington, *supra* note 6, at 45.

[225] *See* Ohlhausen, *supra* note 12.

5. The user technologically **"locks" the algorithm** so that it is difficult to change it. This creates a long-term commitment, or a credible threat that can strengthen coordination, generally without a procompetitive justification.

In all these cases, firms communicate their intentions to act in a certain way, as well as their reliance on one another to follow suit. They do so by using avoidable acts that lack a competitive rationale but that facilitate coordination. Acts that fall under any of these categories in markets, where supra-competitive parallel pricing is observed, should raise red flags and trigger a deeper investigation into procompetitive justifications. The remedy is clear and easy to apply. Of course, when only one side takes action, the conduct might not amount to an agreement in restraint of trade, but rather to an attempt for such an agreement.

Enforcement is likely to become an up-hill battle. Indeed, as the Google Brain experiment noted above indicates, detection and enforcement will become much harder once algorithms autonomously encrypt their messages.[226] Accordingly, antitrust authorities may need to strengthen their technological expertise by employing regulatory algorithms or computer scientists. Nonetheless, several features of algorithms may make such regulatory tasks easier. Algorithms' decision trees reveal the considerations taken into account in reaching decisions.[227] Moreover, algorithms can be tested by running them on specific data, thereby indirectly exposing their decisional parameters.[228] Finally, algorithms can be used by regulators to police and understand the operations of other algorithms.[229] For example, they can be used to determine whether, absent transparency of one's competitors' algorithms, the market equilibrium would have been set at such a high level. By using their resources, authorities can further identify cases which raise red flags, which are based on understanding how algorithms work in the digital environment.

B. THE WAY FORWARD: WIDENING THE NET

The above discussion remains within the confound of existing conceptions of "agreement." While it explored the width of existing laws to capture some

[226] FTC Commissioner McSweeny also recognized the increased detection challenges created by algorithms. Terrell McSweeny, Former Comm'r, Fed. Trade Comm'n, Remarks at Univerisity of Oxford Center for Competition Law and Policy: Algorithms and Coordinated Effects (May 22, 2017).

[227] *See Decision Tree*, WIKIPEDIA, https://en.wikipedia.org/wiki/Decision_tree [https://perma.cc/X2NN-KUST].

[228] *See* EZRACHI & STUCKE, *supra* note 5.

[229] *See* Gal, *supra* note 53, at 6.

types of algorithmic-facilitated coordination, using existing laws to deal with algorithmic-facilitated coordination is not a panacea. Most importantly, as Antonio Capobianco and Anita Nvesto from the Organization of Economic Cooperation and Developmemt observed, "[o]ne of the main risks of algorithms is that they expand the grey area between unlawful explicit collusion and lawful tacit collusion, allowing firms to sustain profits above the competitive level more easily without necessarily having to enter into an agreement."[230] Indeed, as the above analysis showed, the risk of increased conscious parallelism, facilitated by algorithms, is likely to increase. While coordination is not inevitable, sustaining such coordination is strengthened by the inherent characteristics of digital markets and by the increased abilities of algorithms, often without a need to recourse to formal communication or agreement. The use of an algorithm to solve a complex joint profit-maximizing objective that will produce immediate results, and could be followed by others in the market, might not be captured under existing laws. More fundamentally, the fact that algorithms act as "recipes for action" create a situation that is likened to explicit communication.[231] Yet the fact that the algorithm can sometimes be observed indirectly (through reverse-engineering of its actions), limits the ability to capture it under current prohibitions.

Accordingly, unless we treat every algorithm that helps facilitate coordination as a plus factor—a suggestion which is highly problematic—current interpretation of the term "agreement" is likely to leave out many welfare-reducing instances. While the use of autonomous algorithmic interactions to set trade terms has not yet become mainstream, firms have strong incentives to do so. If algorithms can determine trade terms better than humans, and the resulting coordination might be considered legal, there is a strong motivation to use them.

Accordingly, there is an urgent need for a renewed discussion of whether and how current laws should be changed to fit a world that has dispensed with the need for meetings, conversations, and price announcements. The importance of such an analysis is based on the findings of this Article. First, instances of coordination through algorithms are likely to become more commonplace in our digital world.[232] This also implies that one of the considerations underlying the rule which treats conscious parallelism as legal—that it can take place only in a limited number of highly concentrated markets and is therefore likely to create minor economic effects—no longer holds.

[230] Antonio Capobianco & Anita Nyeso, *Challenges for Competition Law Enforcement and Policy in the Digital Economy*, 9 J. EUROPEAN COMPETITION L. & PRAC. 19, 25 (2017).

[231] Salcedo, *supra* note 5; Schwalbe, *supra* note 5, at 16.

[232] *See supra* Part II.

Second, current rules were designed to fit a world characterized by inherent limitations on the human capacity to reach coordination.[233] As the digital world increasingly overcomes these limitations, making it easier to reach agreements, monitor compliance, and apply immediate sanctions, the law will axiomatically capture fewer instances of coordination than it did before. Furthermore, the digital world increases the "paradox of proof," in that market conditions make it easier to coordinate, and at the same time make it more difficult to prove the existence of an explicit agreement given that explicit interfirm communication may be less essential.[234] This suggests that, while the danger of harm might increase, it might also be less likely to find strong evidentiary inferences of an agreement.[235] It is thus the time to rethink our laws and focus on reducing harms to social welfare rather than on what constitutes an agreement. There may well be a case for not binding ourselves to past formulations which no longer fit economic realities.[236] In particular, the time may be ripe to reconsider prohibiting any conduct with potential anticompetitive tendencies with no offsetting pro-competitive ones, even where such conduct does not constitute an agreement in the traditional sense.

V. CONCLUSION

The new world in which algorithms make many business decisions challenges some of our most basic assumptions about how markets operate. As shown, algorithms can make coordination easier and quicker than ever, thereby reducing incentives to compete. This in turn, increases the importance of tools to curtail potential welfare-reducing effects, while ensuring that consumers can enjoy the benefits offered by the digital world. This Article explored some of the challenges to competition created by algorithms used by competitors, as well as some potential market-based and legal countermeasures. In particular, it explored the application of the legal constructs of facilitating practices and plus factors to algorithms, and it suggested a subset of cases which fall under existing rules. As shown, existing laws can capture some of the cases in which algorithms facilitate coordination, yet significant challenges remain.

We are already playing catch-up with technological developments in the use of algorithms and will likely continue to do so. But given the welfare stakes

[233] *See supra* Part III.

[234] *See* KAPLOW, *supra* note 20, at 124–73.

[235] *See id.* at 305.

[236] One such interesting suggestion was made by Harrington, *supra* note 6, at 48–49 (suggesting that some types of pricing algorithms that support supra-competitive prices be per se prohibited, such as reinforcement learning price setting algorithms).

involved, our only option is to brace ourselves for the road ahead and make sure we are as prepared as possible. As one court noted, "the advancement of technological means for the orchestration of large-scale price-fixing conspiracies need not leave antitrust law behind."[237] This Article takes a step in this direction.

[237] Spencer Meyer et al., v. Travis Kalanik, 2016 WL 1266801 15 Civ. 9796 (District Court, S.D. New York, March 31, 2016), Section 7.

CLOUD INFRASTRUCTURE-AS-A-SERVICE AS AN ESSENTIAL FACILITY: MARKET STRUCTURE, COMPETITION, AND THE NEED FOR INDUSTRY AND REGULATORY SOLUTIONS

Kamila Benzina[†]

ABSTRACT

This Note examines whether public cloud infrastructure-as-a-service (IaaS) has a market structure that incentivizes a small number of cloud providers to engage in anticompetitive conduct to the detriment of competitors, competition, and ultimately consumers. As cloud IaaS becomes the dominant model for configuring and delivering computing resources in our increasingly cloud-based economy, the U.S. IaaS market is consolidating around a small number of players. These dominant players—Amazon, Microsoft and Google—also have a significant presence in downstream markets, which creates strong incentives for these providers to leverage their IaaS market power to distort competition in the diverse markets that depend on access to IaaS. While there is the potential for IaaS providers to *act* anticompetitively, the larger challenge is a structural one—*ineffective competition*, which results in a market structure that incentivizes anticompetitive conduct. Given the increasingly vital role cloud IaaS plays in our economy, as well as in our connected lives, important questions emerge as to whether national regulators should take steps to ensure consumers and competition are protected in the emerging cloud-based economy. This Note gives an overview of the IaaS market and examines whether the cost structure of the market has facilitated, and will continue to facilitate, the dominance of a small number of IaaS providers. It goes on to explore how consolidated control over IaaS incentivizes conduct that is potentially harmful to consumers and competitors in varied other markets that depend on access to IaaS. This Note finally explores possible industry and regulatory solutions for ensuring consumers and competition are protected in the emerging cloud-based economy.

DOI: https://doi.org/10.15779/Z38QV3C43D

[†] J.D. Candidate, 2019, Georgetown University Law Center. Thank you to Professor Scott Hempling for his guidance throughout the writing process.

TABLE OF CONTENTS

I. INTRODUCTION

"Friends don't let friends build data centers," proclaimed Infor CEO Charles Phillips, when he announced global software company Infor would shift to a "cloud-first" development approach and move all of its IT operations onto Amazon's cloud computing platform.[1] The motto aptly captures the profound paradigm shift taking place in the way foundational computing resources are configured and delivered in our Internet-dependent economy—a shift to what may be understood as "utility computing."[2]

The National Institute of Standards and Technology (NIST) defines cloud computing as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources . . . that can be rapidly provisioned and released with minimal management effort or service provider interaction."[3] In simpler terms, "cloud computing is on-demand access to virtualized IT resources that are housed outside of your own data center, shared by others, simple to use, paid for via subscription, and accessed over the Web."[4]

Cloud computing is an umbrella concept, encompassing a multilayered ecosystem of IT services that connect users to a variety of resources through

[1] Kate Miller, *Friends Don't Let Friends Build Data Centers*, AWS PARTNER NETWORK BLOG (Mar. 15 2016), https://aws.amazon.com/blogs/apn/friends-dont-let-friends-build-data-centers/ [https://perma.cc/92X2-A3C8].

[2] *See* Bob O'Donnell, *Cloud Computing As a Utility Is Going Mainstream*, RECODE (Aug. 17, 2016), https://www.recode.net/2016/8/17/12519046/cloud-computing-as-utility-private-public-data-center [https://perma.cc/USV2-GDE3].

[3] PETER MELL & TIMOTHY GRANCE, THE NIST DEFINITION OF CLOUD COMPUTING: RECOMMENDATIONS OF THE NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY 2 (2011), http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf [https://perma.cc/2UJH-RMJV].

[4] Erik Brynjolfsson et al., *Cloud Computing and Electricity: Beyond the Utility Model*, 53 COMM. ACM 32, 34 (2010).

web-based tools and applications. This paper will focus on the foundational layer in that ecosystem, cloud infrastructure-as-a-service. Infrastructure-as-a-service is an automated offering where hardware resources, such as servers, networking, hard drive space, and operating systems, are owned and hosted by a service provider and offered to customers on a pay-per-use consumption model. [5] Instead of building expensive IT infrastructure on-premises, businesses large and small can purchase access to scalable, expert-managed and hosted IT services. The cloud computing model is paving the way for tomorrow's cutting-edge services and products that rely on large amounts of data—driverless cars, artificial intelligence, and everything encompassed in the Internet of Things, to name a few.[6]

In the rapidly expanding cloud services market, the infrastructure-as-a-service market is expected to show the fastest growth among the various cloud offerings over the next three years.[7] As prices for infrastructure-as-a-service go down and the U.S. market for cloud infrastructure service consolidates around a handful of big players, small and medium size infrastructure-as-a-service providers are finding it hard to compete with the largest providers. While there is the potential for infrastructure-as-a-service providers to *act* anticompetitively, the larger challenge is a structural one—*ineffective competition*, which results in a market structure that incentivizes anticompetitive conduct. Given the increasingly vital role cloud infrastructure-as-a-service plays in our economy, as well as in our connected lives, important questions emerge as to whether national regulators should take steps to ensure consumers and competition are protected in the emerging cloud-based economy.

The purpose of this Note is to examine whether public cloud infrastructure-as-a-service has a market structure that incentivizes a small number of cloud providers to engage in anticompetitive conduct, to the detriment of competitors, competition, and ultimately consumers. Part II gives an overview of the infrastructure-as-a-service market. Part III discusses the cost structure of the market and how this cost structure has facilitated, and will continue to facilitate, the dominance of a small number of providers. Part IV goes on to discuss why consolidated control over this essential facility incentivizes conduct that is potentially harmful to consumers and competitors. Lastly, Part V explores possible industry and regulatory solutions for a future where connectivity and cloud computing are ubiquitous.

[5] *See* MELL & GRANCE, *supra* note 3, at 2–3.

[6] *See* Mike Chan, *Why Cloud Computing Is the Foundation of the Internet of Things*, THORN TECHS. (Feb. 15, 2017), https://www.thorntech.com/2017/02/cloud-computing-foundation-internet-things/ [https://perma.cc/5XSX-MYV4].

[7] *See* Press Release, Gartner, Inc., Gartner Forecasts Worldwide Public Cloud Revenue to Grow 21.4 Percent in 2018 (Apr. 12, 2018) [hereinafter Gartner Press Release on Forecast].

II. OVERVIEW OF INFRASTRUCTURE-AS-A-SERVICE MARKET

Infrastructure-as-a-service provides consumers with "provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications."[8] The infrastructure-as-a-service provider hosts and manages the underlying infrastructure, while the cloud user purchases computing resources and data storage as needed, avoiding the expense of building and maintaining on-premises data-center infrastructure.[9] This Section gives an overview of the infrastructure-as-a-service market, identifying "who" is selling "what" to "whom" in the public cloud infrastructure-as-a-service market.

A. WHAT ARE INFRASTRUCTURE-AS-A-SERVICE (IAAS), PLATFORM-AS-A-SERVICE (PAAS), AND SOFTWARE-AS-A-SERVICE (SAAS)?

In *public* cloud infrastructure-as-a-service (IaaS), providers create a shareable multi-tenant IT infrastructure by "virtualizing" hardware resources, using specialized software to break hardware into discrete and separate units, for purchase on a pay per use basis.[10] Key to the public cloud IaaS model is that the provider's resources are pooled to serve a multitude of consumers, dynamically provisioned and deprovisioned according to the demands of individual consumers.[11] In a *private* cloud model, computing resources are provisioned for exclusive use by a single organization, eliminating most of the cost benefits gained from resource pooling in the public model.[12]

The development and deployment of other cloud services are broadly categorized as platform-as-a-service (PaaS) and software-as-a-service (SaaS). The three layers of the cloud computing stack—IaaS, PaaS, and SaaS—encompass three different but interdependent services, with most SaaS and PaaS running atop IaaS.

[8] MELL & GRANCE, *supra* note 3, at 3.

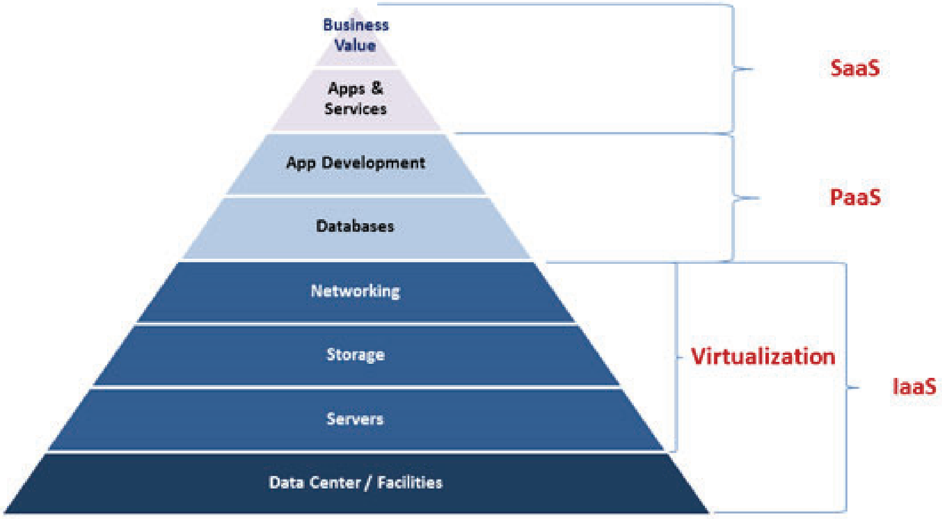
[9] *See id.*

[10] *See* Sreedhar Kajeepeta, *Multi-tenancy in the cloud: Why It Matters*, COMPUTERWORLD (Apr. 12, 2017), https://www.computerworld.com/article/2517005/multi-tenancy-in-the-cloud--why-it-matters.html [https://perma.cc/38AY-V925]; MELL & GRANCE, *supra* note 3, at 2 n.1.

[11] *See* MELL & GRANCE, *supra* note 3, at 2–3.

[12] *See id.* at 3.

Figure 1: Hierarchy of Cloud Offerings[13]



Platform-as-a-service (PaaS) allows consumers "to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages, libraries, services, and tools supported by the provider."[14] PaaS provides additional functionality on top of IaaS, allowing developers to host, build, test, and deploy applications without having to manage the underlying computing resources.[15] An example of cloud PaaS is Amazon Elastic Beanstalk, a platform that makes it easier for users to deploy and manage applications in the AWS Cloud by providing functions such as load balancing and application health monitoring.[16]

Software-as-a-service (SaaS) is the capability provided to the consumer "to use the provider's applications running on a cloud infrastructure."[17] SaaS is the "[delivery of] applications that are managed by a third-party vendor and whose interface is accessed on the clients' side."[18] Most SaaS applications can be run

[13] *See* Ephraim Baron, *Aren't Virtualization and Cloud the Same Thing?*, EQUINIX (Nov. 2, 2011), https://blog.equinix.com/blog/2011/11/02/aren%E2%80%99t-virtualization-and-cloud-the-same-thing [https://perma.cc/U4BK-Y323].

[14] MELL & GRANCE, *supra* note 3, at 2–3.

[15] *See IaaS, PaaS, SaaS (Explained and Compared)*, APPRENDA, https://apprenda.com/library/paas/iaas-paas-saas-explained-compared/ [https://perma.cc/XR67-ANE7] (last visited Apr. 2, 2019) [hereinafter *IaaS, PaaS, SaaS (Explained and Compared)*].

[16] *See* AWS ELASTIC BEANSTALK, https://aws.amazon.com/elasticbeanstalk/ [https://perma.cc/2YH9-3TR5] (last visited Apr. 2, 2019).

[17] MELL & GRANCE, *supra* note 3, at 2.

[18] *IaaS, PaaS, SaaS (Explained and Compared)*, *supra* note 15.

directly from a web browser and do not have to be downloaded. The average Internet user likely encounters SaaS multiple times a day by using cloud-based applications such as Google Apps, Microsoft 365, or Netflix. SaaS has also emerged as a major component of the IT systems of large companies and organizations, allowing these entities to purchase and rent software licenses as needed with the flexibility to scale up and down as their needs evolve.[19] Of the three service layers, the SaaS market is the largest, with an array of providers offering varying platforms and applications.[20]

B. WHO IS BUYING PUBLIC CLOUD IAAS?

To illuminate who is buying public cloud IaaS, it is helpful to identify the most common use cases of IaaS and then discuss the various types of customers of the service.

1. *What Are the Use Cases of Public Cloud IaaS?*

To illuminate *who* is buying public cloud IaaS, it is helpful to first identify the four most common use cases of IaaS: (1) development and testing, (2) website hosting, (3) enterprise applications, and (4) cloud native applications.[21] Many developers turn to IaaS as a cost-effective way to build, test, and deploy applications. [22] Unlike traditional, on-premises infrastructure, IaaS allows developers to rapidly self-provision testing environments and scale up and down as needed. Developers realize cost savings and time efficiencies when using cloud-based development environments, rather than building infrastructure or tapping into existing on-premises infrastructure.[23]

Individuals and organizations also use public cloud IaaS for hosting Internet-facing websites and web-based applications, such as SaaS. These are the websites and applications that the everyday user interacts with online. Unlike traditional hosting agreements, where a user pays a flat fee for a set amount of storage and processing power, IaaS provides resources on an on-

[19] *See* Sebastian Lambert, *2018 SaaS Industry Market Report: Key Global Trends & Growth Forecasts*, FINANCESONLINE, https://financesonline.com/2018-saas-industry-market-report-key-global-trends-growth-forecasts/ [https://perma.cc/6RKM-ARHR] (last visited Apr. 2, 2019).

[20] *See* Gartner Press Release on Forecast, *supra* note 7.

[21] *See* LYDIA LEONG ET AL., MAGIC QUADRANT FOR CLOUD INFRASTRUCTURE AS A SERVICE, WORLDWIDE (June 15, 2017).

[22] *See* David Linthicum, *Why Application Development Is Better in the Cloud*, INFOWORLD (Jan. 25, 2013), https://www.infoworld.com/article/2613509/paas/why-application-development-is-better-in-the-cloud.html [https://perma.cc/U94P-HBRL].

[23] *See id.*

demand, as-needed basis.[24] When necessary, the provider can balance the information load across a number of servers in a cluster configuration to accommodate wide variation in usage, scaling up and down automatically with website traffic.[25] IaaS providers, such as Amazon Web Services, offer cloud web hosting solutions to customers looking to host a variety of websites on their infrastructure, from small-scale "simple website hosting" to large scale "enterprise web hosting," for either fixed monthly fees or pay-as-you-go pricing plans.[26]

Enterprise applications are general-purpose workloads used by businesses internally to perform various business functions.[27] These kinds of workloads could traditionally be found in the "on-premises" (company owned and operated) data centers of enterprise IT environments, including storage and operating systems. Enterprise applications range from automated billing systems to enterprise content management, and they demand reliable and high-performance infrastructure.[28]

Lastly, there are cloud native applications, which are specifically architected to run in a cloud infrastructure environment.[29] Cloud native applications are designed from the ground-up on the cloud (rather than being migrated to the cloud from an on-premise data center), allowing the application developers to exploit innovative ways of designing, partitioning, scaling, testing and deploying. These include applications in the rapidly emerging Internet of Things, which require high availability, flexibility, and scalable capacity.[30]

[24] *See Cloud Hosting vs. Traditional Hosting*, OPUS:INTERACTIVE, http://www.opusinteractive.com/cloud-hosting-vs-traditional-hosting/ [https://perma.cc/ZK2X-6SVQ] (last visited Apr. 2, 2019).

[25] *See id.*

[26] *Web Hosting*, AMAZON WEB SERVICES, https://aws.amazon.com/websites/ [https://perma.cc/J8J9-EAPZ] (last visited Apr. 2, 2019).

[27] *See* Patrick Hogan, *Why Infrastructure as a Service (IaaS) Works for Enterprise-Level Companies*, TENFOLD, https://www.tenfold.com/iaas/iaas-enterprise-companies [https://perma.cc/2AXN-53FR] (last visited Apr. 2, 2019).

[28] *See id.*

[29] *See* Rishi Yadav, *What Real Cloud-Native Apps Will Look Like*, TECHCRUNCH (Aug. 3, 2016), https://techcrunch.com/2016/08/03/what-real-cloud-native-apps-will-look-like/ [https://perma.cc/B7M3-Q3RQ].

[30] *See* Andrew Meola, *The Roles of Cloud Computing and Fog Computing in the Internet of Things Revolution*, BUSINESS INSIDER (Dec. 20, 2016), https://www.businessinsider.com/internet-of-things-cloud-computing-2016-10 [https://perma.cc/CPP7-L38Z].

2. *Who Are the Main Customers of Public Cloud IaaS?*

Whereas SaaS and PaaS are commonly employed by a range of consumers, IaaS is directed primarily toward enterprise-level companies, organizations, and government entities that operate on a large scale.[31] IaaS gives companies control over IT, but requires extensive expertise on the part of the customer to manage the computing infrastructure.[32] Most small to medium-size businesses opt for PaaS and SaaS solutions that allow them to use cloud-based applications without needing to manage the underlying infrastructure.[33] Providers of these PaaS and SaaS solutions are increasingly turning to the large IaaS providers, such as Amazon Web Services, rather than attempting to continue competing in the infrastructure-as-a-service market.[34]

One 2016 study predicted over half of all large enterprises will adopt infrastructure-as-a-service as the primary environment for workloads by 2018.[35] This is not surprising, given that today the average enterprise deploys 464 custom applications, all of which depend on the underlying infrastructure that the large IaaS providers are selling at increasingly lower prices.[36] Many major companies are already running operations fully through cloud-based services, while others are integrating cloud IaaS into their IT systems incrementally. In 2017, 75% of hospitals' Chief Information Officers planned to use IaaS within a year—up from just 15.3% in 2014.[37] Higher education institutions are also adopting cloud-first policies for running on-campus IT infrastructure and are heavy consumers of public cloud IaaS.[38]

The other major customers of IaaS are federal, state, and local government entities. In 2016, the federal government spent nearly $8.5 billion on cloud

[31] *See* Hogan, *supra* note 27.

[32] *See* Gleb B., *Choosing the Right Cloud Service: IaaS, PaaS, or SaaS*, RUBYGARAGE https://rubygarage.org/blog/iaas-vs-paas-vs-saas [https://perma.cc/M7U9-J4LW] (last visited Apr. 2, 2019).

[33] *See* Leong et al., *supra* note 21.

[34] *See id.*

[35] *See* Arul Elumalai et al., *IT as a Service: From Build to Consume*, MCKINSEY & COMPANY (Sept. 2016), https://www.mckinsey.com/industries/high-tech/our-insights/it-as-a-service-from-build-to-consume [https://perma.cc/PU3F-9B4L].

[36] *See* SKYHIGH NETWORKS, CUSTOM APPLICATIONS AND IAAS TRENDS 4–5 (2017).

[37] Laleh Hassibi, *Saas, Paas, IaaS; What's the Difference?*, DATICA BLOG (July 4, 2017), https://datica.com/blog/saas-paas-iaas-whats-the-difference/ [https://perma.cc/BE64-WADM].

[38] *See, e.g.*, Brandon Butler, *How Notre Dame Is Going All in with Amazon's Cloud*, NETWORK WORLD (Dec. 14, 2015), https://www.networkworld.com/article/3014599/cloud-computing/how-notre-dame-is-going-all-in-with-amazon-s-cloud.html [https://perma.cc/4SDC-CJUK] (reporting on University of Notre Dame's 2015 adoption of a cloud-first policy).

IaaS, leveraging both public and private clouds.[39] In response to the federal government's "Cloud First" policy (now the "Cloud Smart" policy),[40] Amazon Web Services launched GovCloud, an isolated data center region which exclusively serves federal government entities and partners.[41] The GovCloud adheres to strict security standards under the Federal Risk and Authorization Management Program that Amazon's public cloud does not necessarily meet.[42] GovCloud still leverages the resource pooling and multi-tenant benefits of a public cloud, but limits access to a specific community of customers. Amazon's GovCloud clients include NASA, the United States Airforce, and the Department of Justice, as well as US government contractors like Lockheed Martin.[43] Microsoft, Google, and IBM have been authorized to provide similar cloud offerings to the federal government and are eager to compete for dominance in the federal cloud market.[44]

C. WHO IS SELLING PUBLIC CLOUD IAAS?

Amazon was the first to arrive at the IaaS market in 2006,[45] and today commands more than half of the $23.5 billion global cloud IaaS market.[46] The top five IaaS providers—Amazon, Microsoft, Alibaba, Google, and IBM—control 75% of the global IaaS market, with Amazon and Microsoft dominant in the United States. In 2017, Amazon Web Services had a 51.8% share of the

[39] *The Bumps, Cuts and Zeros in Trump's Tech Budget*, NEXTGOV (May 23, 2017), http://www.nextgov.com/cio-briefing/2017/05/bumps-cuts-and-zeros-trumps-tech-budget/138096/ [https://perma.cc/49T5-JYAJ].

[40] *See Cloud Smart Strategy*, U.S. DEP'T OF THE INTERIOR, https://www.doi.gov/cloud/strategy [https://perma.cc/5TCE-PHNY] (last visited Apr. 2, 2019).

[41] *See Introduction to the AWS GovCloud (US) Region*, AMAZON WEB SERVICES, https://aws.amazon.com/govcloud-us/ [https://perma.cc/9UEF-CXK2] (last visited Apr. 2, 2019).

[42] *Cf. id.*

[43] *See* Cassandra Stephenson, *AWS GovCloud Announces Eastern Expansion*, FEDSCOOP (June 13, 2017), https://www.fedscoop.com/aws-govcloud-announces-eastern-expansion/ [https://perma.cc/M8NG-HHJQ]; *Lockheed Martin Case Study*, AMAZON WEB SERVICES, https://aws.amazon.com/solutions/case-studies/Lockheed-martin/ [https://perma.cc/3TQ5-DSHD] (last visited Apr. 2, 2019).

[44] Frank Konkel, *Google Cloud Targets Federal Government*, NEXTGOV (Mar. 23, 2018), http://www.nextgov.com/it-modernization/2018/03/google-cloud-targets-federal-government/146917/ [https://perma.cc/L789-4BNQ].

[45] *See* Ron Miller, *How AWS Came to Be*, TECHCRUNCH (July 2, 2016), https://techcrunch.com/2016/07/02/andy-jassys-brief-history-of-the-genesis-of-aws/ [https://perma.cc/9YTN-LR3K].

[46] *See* Press Release, Gartner, Inc., Gartner Says Worldwide IaaS Public Cloud Services Market Grew 29.5 Percent in 2017 (Aug. 1, 2018) [hereinafter Gartner Press Release on Growth].

global cloud IaaS market.[47] Its next biggest competitor was Microsoft Azure, which had 13.3% market share.[48] Together, Amazon and Microsoft represented 75% of IaaS industry growth in 2017—Amazon with 45% and Microsoft 28.9%.[49]

The next biggest competitor is Chinese e-commerce giant Alibaba, which commands 4.6 % of the global market share,[50] garnering most of its business in Asia.[51] U.S. companies Google and IBM trail with 3.3% and 1.9% of the IaaS market share, respectively.[52] The remainder of the IaaS market is highly fragmented, with small and mid-sized cloud providers facing decreasing market shares.[53] As competing providers are increasingly unable to meet the prices offered by major providers such as Amazon and Microsoft, they are moving away from hosting their own infrastructure, opting instead to help companies manage and implement their use of the major players' IaaS.[54]

As Gartner's analysis highlights, "[c]loud IaaS providers have increasingly openly acknowledged that they cannot compete directly against the market leaders for public cloud IaaS. Many such providers that historically have managed hosting businesses have pivoted to offer their managed services on top of market-leading cloud IaaS platforms instead."[55] While Google, IBM, and Alibaba's IaaS revenues grew substantially in 2017,[56] these three companies, along with all other IaaS providers combined, only represented 25% of overall IaaS industry growth,[57] which suggests the two largest IaaS providers—Amazon and Microsoft—are not ceding ground in the near future.

[47] *See id.*

[48] *Id.*

[49] Jamal Carnette, *Microsoft Is Taking On Amazon's Profit Center*, MOTLEY FOOL (Aug. 13, 2018), https://www.fool.com/investing/2018/08/13/microsoft-is-taking-on-amazons-profit-center.aspx [https://perma.cc/SE5R-Q46V].

[50] *See* Gartner Press Release on Growth, *supra* note 46.

[51] *See* Ron Miller, *Alibaba Continues to Gain Cloud Momentum*, TECHCRUNCH (Aug. 24, 2018), https://techcrunch.com/2018/08/24/alibaba-continues-to-gain-cloud-momentum/ [https://perma.cc/D5SU-Z7CP].

[52] *See* Gartner Press Release on Growth, *supra* note 46.

[53] *See* Leong et al., *supra* note 21.

[54] *See id.*

[55] *Id.*

[56] *See* Gartner Press Release on Growth, *supra* note 46.

[57] *See* Miller, *supra* note 51.

III. IS EFFECTIVE COMPETITION POSSIBLE IN THE IAAS MARKET?

The barriers to entry that exist in the U.S. IaaS market make it difficult for new entrants and smaller IaaS providers to compete with Amazon, Microsoft, and Google. A "barrier to entry" is "[a]ny market condition that makes entry more costly or time-consuming and thus reduces the effectiveness of potential competition as a constraint on the pricing behavior of the dominant firm . . . ."[58] The major barriers to entry for new competitors in the IaaS market include: the incumbents' large sunk costs; technological leadership and reputation; and customer switching cost and inconvenience. The following Sections present these barriers to entry in more detail and discuss how they are already facilitating concentration in the IaaS market.

A. THE LARGEST INCUMBENT IAAS PROVIDERS BENEFIT FROM ECONOMIES OF SCALE, WHILE NEW ENTRANTS ARE UNABLE TO COMPETE DUE TO LARGE SUNK COSTS

"Economies of scale" is "the phenomenon where the average costs per unit of output decrease with the increase in the scale or magnitude of the output being produced by a firm."[59] As public cloud IaaS providers increase output, input costs per unit decrease because of bulk buying, organizational efficiencies, and virtualization technology that allows for high utilization of data infrastructure.

To enter the IaaS market, a firm must make enormous investments in facilities and the equipment, as well as in the technology that will allow for essential cloud characteristics such as high elasticity and scalability. IaaS providers must build and maintain data centers that operate twenty-four hours a day, seven days a week. Large data centers contain tens of thousands of servers that require reliable and continuous power, cooling, and connectivity.[60] Smaller providers are not able to achieve this scale without massive investment in physical infrastructure and hiring expert teams to configure and manage the infrastructure. An IaaS provider must buy the expensive hardware that goes in the datacenters, the physical space needed to house the servers, and the utility services that will keep the datacenter running. One study showed "that very

[58] S. Pac. Commc'ns Co. v. AT&T, 740 F.2d 980, 1001 (D.C. Cir. 1984).

[59] *Glossary of Statistical Terms*, THE ORG. FOR ECON. CO-OPERATION & DEV., https://stats.oecd.org/glossary/detail.asp?ID=3203 [https://perma.cc/F5PN-69QC] (last visited Apr. 2, 2019).

[60] Michele Lerner, *Data Centers and the Cloud Are Going Green*, NAREIT (July 20, 2016) https://www.reit.com/news/reit-magazine/july-august-2016/data-centers-and-cloud-are-going-green [https://perma.cc/3SWE-9PD4].

large datacenters (tens of thousands of computers) can purchase hardware, network bandwidth, and power for 1/5 to 1/7 the prices offered to a medium-sized (hundreds or thousands of computers) datacenter."[61] Entry to the IaaS market requires a large amount of capital and those entrants who can invest very large amounts have an automatic advantage in terms of input cost.

B. The Importance of Technological Leadership in Achieving Economies of Scale Makes It Difficult for New Entrants or Smaller Providers to Compete

Leading in the cloud IaaS space is not just about the size of a company's data centers. Leveraging the economies of scale offered by public cloud IaaS requires innovative virtualization, automation, dynamic scaling, and metering technology that allow for both higher utilization of data infrastructure and the spreading of fixed hardware and software costs over many more machines, and therefore over many more consumers.[62] These technologies enable IaaS providers to pool resources to dynamically serve multiple consumers with varying demands for physical and virtual resources.[63]

This is the way IaaS companies will continue to differentiate their products and prices, making it difficult for smaller providers to compete with those companies that are leaders in developing and patenting new cloud technology. The secret to Amazon Web Service's success has been the evolution of its datacenter technology, which has advanced over time to allow for higher utilization of its networks.[64] While Amazon Web Services entered the IaaS market in 2006 by purchasing basic servers, the company realized the benefits of developing customized technology that allowed for more sophisticated uses of IaaS.[65] The datacenters have become more advanced, and the systems have evolved as well, allowing for higher utilization of networks.

IT giants like Amazon, Google, and Microsoft have the money to put into research and patent licensing over these technologies that other providers do not.[66] For example, Amazon has patents claiming technology that removes the

[61] Michael Armbrust et al., Above the Clouds: A Berkeley View of Cloud Computing 5 (2009).

[62] *See id.*

[63] *See* Timothy Prickett Morgan, *A Rare Peek into The Massive Scale of AWS*, Enterprise AI (Nov. 14, 2014), https://www.enterpriseai.news/2014/11/14/rare-peek-massive-scale-aws/ [https://perma.cc/CYJ4-U8ZF].

[64] *See id.*

[65] *See id.*

[66] *See Winners And Losers In The Patent Wars Between Amazon, Google, Facebook, Apple, and Microsoft*, CB Insights (Nov. 16, 2017), https://www.cbinsights.com/research/innovation-patents-apple-google-amazon-facebook-expert-intelligence/ [https://perma.cc/59VW-

complexity associated with provisioning, administering, and managing resources of data centers, as well as technology that monitors the execution of a computing service to reduce errors and delays.[67] This patented technology enables Amazon Web Services to improve management and monitoring, which are key to optimizing utilization of infrastructure resources.

C. NEW ENTRANTS AND SMALL IAAS PROVIDERS FACE REPUTATIONAL BARRIERS TO ENTRY TO THE IAAS MARKET AND LACK THE MARKET KNOWLEDGE THAT BREEDS CUSTOMER LOYALTY

Amazon, Microsoft, and Google are global technology conglomerates that span industries from cloud computing (IaaS, PaaS, SaaS) to search engines to online retail. With high name recognition and a broad array of customer connections, these IaaS providers do not require as much capital to attract IaaS customers. Their reputation allows for lower per-customer attraction costs and facilitates customer loyalty.

In the case of Amazon—a company that was traditionally a leader in online retail—many of Amazon Web Service's loyal customers in the IaaS space come from its earliest days as one of the only IaaS providers. Most of these early customers were small developers who used Amazon Web Service's infrastructure offerings as a cheap way to test or run simple websites.[68] These small startups include the likes of what are now Netflix and Airbnb, who still rely on Amazon Web Services for their cloud infrastructure needs.[69]

The leading IaaS providers also have long-term working relationships in other markets which give them cheaper access to customers. For example, Microsoft's dominance in software—driven by its most popular SaaS offering, Office 365—has given it an established customer base and deep knowledge of enterprise IT.[70] Time and money spent in the IaaS market plus well-established data analytics teams mean the major providers also have more information about customers, such as their various needs and uses for IaaS, as well as their consumption patterns. Market knowledge, customer knowledge, and internal

FAUM]; *Patent Transaction Trends in Cloud Computing: Are Paes Buying Into The Market?*, IAM-MEDIA (Oct. 25, 2017), https://www.iam-media.com/patent-transaction-trends-cloud-computing-are-paes-buying-market [https://perma.cc/7P53-A9TY].

[67] Steve Brachmann, *A Modest Patent Portfolio Doesn't Stop Amazon Web Services from Earning $5.16 Billion*, IPWATCHDOG (May 1, 2015), http://www.ipwatchdog.com/2015/05/01/modest-patent-portfolio-amazon-web-services/id=57252/ [https://perma.cc/8W52-GZMM].

[68] *See* Matt Weinberger, *The Cloud Wars Explained: Amazon Is Dominating, but Microsoft and Google Are Striking Back*, BUSINESS INSIDER (July 22, 2017), https://www.businessinsider.com/why-amazon-is-so-hard-to-topple-in-the-cloud-and-where-everybody-else-falls-2017-7 [https://perma.cc/7P5Y-94H6].

[69] *See id.*

[70] *See id.*

expertise also breed customer loyalty, meaning existing customer relationships are more readily (less costly) maintained.

D. The Costs and Security Risks of Switching IaaS Providers Makes It Difficult for New Entrants or Providers with Small Market Shares to Compete for Customers Already "Locked-In" to Another IaaS Provider

Once entities move to public cloud IaaS, it is not easy to switch providers. One survey showed the majority of businesses interviewed would not change IaaS providers even for a 20% discount, as switching would bring new risks and added costs in the re-development of tools on the new provider's interface.[71] High switching costs and risk disadvantages new entrants who are seeking to lure IaaS customers away from the major providers. Because it is difficult and costly to port personal and business data from one infrastructure provider to another, users may end up locked in to a suboptimal contractual relationship, with no feasible way to switch to competing offers.

The continued rapid growth of Amazon, Microsoft, and Google, as well as the departure of other smaller IaaS providers, suggests the market is moving toward an oligopoly, if not a duopoly. These barriers to entry make it more likely that a handful of major providers will make it difficult for smaller existing IaaS providers or new entrants to compete, thereby maintaining and continuing to expand their market power in an already concentrated IaaS market.

IV. WHY IS INEFFECTIVE COMPETITION IN THE IAAS MARKET BAD FOR COMPETITION AND CONSUMERS?

Increased concentration in the U.S. IaaS market poses several potential threats to consumers. As a leading researcher at Gartner aptly warned, "[t]he increasing dominance of the hyperscale IaaS providers creates both enormous opportunities and challenges for end users and other market participants;" and "[w]hile it enables efficiencies and cost benefits, organizations need to be cautious about IaaS providers potentially gaining unchecked influence over customers and the market."[72]

This Part takes a closer look at the potential threat an IaaS duopoly, or oligopoly, poses to competition in the diverse markets that depend on access to IaaS. IaaS is an essential service, providing the opportunity for a monopoly

[71] *See* John DeWolf, *The Future of the Cloud: Will AWS Continue to Dominate?*, BACKUPIFY (May 9, 2013), https://www.backupify.com/blog/the-future-of-the-cloud-will-aws-continue-to-dominate [https://perma.cc/Q9PS-YAVB].

[72] Gartner Press Release on Forecast, *supra* note 7.

firm, or oligopoly firms, to leverage their market power in IaaS to distort competition in markets that depend on access to IaaS. While it is difficult to prove to what extent this kind of behavior is already taking place, this Part underscores that the incentives to engage in this anticompetitive behavior do exist and will likely become more potent as the cloud market grows and matures.

A. INFRASTRUCTURE-AS-A-SERVICE IS AN ESSENTIAL SERVICE

An essential facility is a facility that (1) is essential for competition and (2) competitors cannot duplicate economically.[73] As described above, the infrastructure layer is the layer on which every other cloud service depends—an application or platform cannot exist without the underlying servers, networks, hard drive space, and operating systems that IaaS provides. Access to IaaS is therefore essential to compete in the PaaS and SaaS markets. According to Gartner, "[b]y 2020, a corporate 'no-cloud' policy will be as rare as a 'no-internet' policy is today." For many of the reasons discussed in Part III, it is becoming increasingly infeasible for most companies, organizations, or individuals to duplicate IaaS in ways that allow them to compete with companies using IaaS provided by AWS, Microsoft, or Google.[74] IaaS is becoming an essential facility for more than just other cloud services. As large companies increasingly depend on cloud computing for their IT operations, access to IaaS also becomes essential to remain competitive in their respective markets.

Some might argue that cloud IaaS is only one way to access the infrastructure layer—many large companies still have on-premises IT or access to a private cloud. However, "[to] be 'essential' a facility need not be indispensable; it is sufficient if duplication of the facility would be economically infeasible and if denial of its use inflicts a severe handicap on potential market entrants."[75] For small and medium size businesses, including emerging cloud-native applications, public cloud services are the only economically feasible way to tap into even the most basic business IT operations; this includes hosting a website, storing large amounts of data, or accessing an e-mail platform.

[73] *See* SCOTT HEMPLING, REGULATING PUBLIC UTILITY PERFORMANCE: THE LAW OF MARKET STRUCTURE, PRICING AND JURISDICTION 135 (2013).

[74] *See* Press Release, Gartner, Inc., Gartner Says By 2020, a Corporate "No-Cloud" Policy Will Be as Rare as a "No-Internet" Policy Is Today (June 22, 2016).

[75] Hecht v. Pro-Football, Inc., 570 F.2d 982, 985, 992 (D.C. Cir. 1977).

B. The Companies That Control IaaS Could Leverage That Market Power to Distort Competition Through Tying and Price Squeezing

As the dominant IaaS providers continue to widen their menu of PaaS and SaaS offerings, they have incentives to leverage their market power in IaaS to distort competition in these IaaS-dependent markets. Other markets vulnerable to such distortion include those that depend on access to IaaS for IT operations and other enterprise applications.

1. Providers with IaaS Market Power Could Distort the SaaS and PaaS Markets by "Tying" Services to IaaS Purchases

IaaS providers could distort markets of services that depend on access to IaaS through "tying": conditioning the sale of their IaaS on the purchase of a different service they offer. For example, Amazon Web Services could require its IaaS platform customers to also purchase Amazon Aurora, a cloud database service that is the company's fastest growing PaaS offering.[76] If a large enterprise customer is already locked into Amazon's IaaS platform, it may be economically infeasible to change providers. In that case, Amazon is effectively forcing their IaaS customers to also purchase their PaaS offerings instead of a competitor's.

Another version of "tying" is "technology tying": where a seller designs a product that only functions when used with that seller's complementary product.[77] In cloud computing, there are a diverse array of interfaces and applications that work together at the IaaS, PaaS, and SaaS levels.[78] In order to exchange information, the different applications and interfaces must be interoperable, where resources on one cloud provider system can communicate with resources on another's cloud provider system.

When it comes to IaaS, interoperability refers to the application programming interfaces "needed so that the virtualization platform's management interfaces to operate between different providers."[79] Amazon Web Services, for example, could configure its infrastructure to enable communication only with its own PaaS and SaaS applications and not others.

[76] *See How Amazon Is Disrupting a $34bn Database Market*, CLOUDTECH (Aug. 11, 2016), https://www.cloudcomputing-news.net/news/2016/aug/11/how-amazon-disrupting-34bn-database-market/ [https://perma.cc/9J67-NFXQ].

[77] *See* HEMPLING, *supra* note 73, at 202.

[78] *See* CLOUD STANDARDS CUSTOMER COUNCIL, INTEROPERABILITY AND PORTABILITY FOR CLOUD COMPUTING: A GUIDE 4 (2017).

[79] Niamh Christina Gleeson & Ian Walden, *'It's a Jungle Out There'?: Cloud Computing, Standards and the Law*, 5 EUR. J.L. & TECH. 461, 462 (2014).

This interoperability is also important to ensure consumers have the ability to migrate workloads between different providers, to prevent customers from being "locked-in" to any one provider because of inaccessibly high costs and risks in migrating data to a new provider's system.

Notably, the major IaaS providers are already selling PaaS offerings that are tightly woven into IaaS offerings. For example, Microsoft Azure currently offers two services: a fully-automated PaaS environment and a do-it-yourself IaaS capability.[80] Microsoft is blurring the lines by releasing certain application extensions that will bring some of the managed functionality of PaaS to the IaaS.[81] Amazon Web Services now also offers more curated services in response to a demand for IaaS that is more user-friendly.[82] Some experts predict that PaaS will be absorbed into IaaS, meaning the myriad features and functions of platform-as-a-service could become systemic to the IaaS platform.[83] This is leading many companies to go "all-in," obtaining all their cloud services from a single provider.[84]

2. *Providers with IaaS Market Power Could "Price Squeeze" to Disadvantage Competitors in Other Markets*

Companies with infrastructure-as-a-service market power could also price squeeze their competitors in other industries that utilize their IaaS for most, or all, of their operations. A "price squeeze" occurs when a vertically-integrated carrier has competitors who depend on an input provided by that carrier.[85] The carrier can harm those "in the downstream market by reducing the margin between the retail price it charges in the downstream market and the wholesale access price it charges [its competitors] for an essential input."[86] For example, Amazon's video streaming service through Amazon Prime competes with

[80] *See* Mary Jo Foley, *Microsoft's Azure Cloud Team Moves Toward Blurring the IaaS/PaaS Lines*, ZDNET (Feb. 24, 2014), http://www.zdnet.com/article/microsofts-azure-cloud-team-moves-toward-blurring-the-iaaspaas-lines/[https://perma.cc/TRK8-YGPC].

[81] *See id.*

[82] *See* Valerie Silverthorne, *AWS Blurs the Lines with PaaS and IaaS*, TECHTARGET (Apr. 2015), http://searchcloudapplications.techtarget.com/tip/AWS-blurs-the-lines-with-PaaS-and-IaaS [https://perma.cc/QJ72-YSUD].

[83] *See* David Linthicum, *Will 2017 Mark the Death of PaaS?*, CLOUD TECH. PARTNERS (Jan. 11, 2017), https://www.cloudtp.com/doppler/paas-death-watch/ [https://perma.cc/RS6Y-6BWX].

[84] *See* Larry Dignan, *Enterprises Learning to Love Cloud Lock-In Too: Is It Different This Time?*, ZDNET (Apr. 8, 2018), https://www.zdnet.com/article/enterprises-learning-to-love-cloud-lock-in-too-is-it-different-this-time/ [https://perma.cc/67VG-29V3].

[85] *Price Regulations – FAQ*, INT'L TELECOMM. UNION, http://www.itu-coe.ofca.gov.hk/vtm/price/faq/q10.htm [https://perma.cc/ZWA9-FDC3] (last visited Apr. 2, 2019).

[86] *Id.*

Netflix, one of Amazon Web Service's most prominent IaaS customers. Because Amazon controls the IaaS that underlies all of Netflix's operations—a service on which Netflix depends—Amazon could disadvantage Netflix by charging a price for the input exceeding the cost to Amazon of self-providing that input, leaving Netflix disadvantaged in competing with Amazon's video streaming prices.

However, competitors are noticing Amazon's control of this powerful resource. After Amazon's recent acquisition of Whole Foods, competitors in the brick-and-mortar supermarket industry, including Target and Wal-Mart, are reportedly scaling back their use of Amazon Web Services, moving e-commerce activities, mobile development, and operations away from Amazon.[87] Walmart is building its own cloud-based data-centers, while Target is looking at using other cloud providers.[88] In fact, one possible explanation for Microsoft's growth in 2017 is that Walmart, Target, Costco, and Walgreens have all opted for Microsoft's off-site servers, storage, and networking services.[89] As described in Part III, however, many companies do not have the resources to build their own infrastructure or go through the expensive process of switching providers, especially as the IaaS market becomes more concentrated and customers have fewer viable options.

3. *Tying and Price Squeezing Could Lead to Reduced Competition in a Wide Range of Industries, Leading to Fewer Choices for Consumers*

Tying can bring both benefits and harm to consumers. On the one hand, bundled offerings of IaaS, PaaS, and SaaS could create efficiencies that lead to lower prices, and benefit consumers that seek the convenience of going to one provider for various services. On the other hand, if the major IaaS providers engage in unlawful tying and price squeezing, they would disadvantage competitors in markets that depend on IaaS. These practices would make it

[87] *See* Christina Farr & Ari Levy, *Target Is Plotting A Big Move Away From AWS as Amazon Takes Over Retail*, CNBC (Aug. 29, 2017), https://www.cnbc.com/2017/08/29/target-is-moving-away-from-aws-after-amazon-bought-whole-foods.html [https://perma.cc/P4TP-QJ24].

[88] *See* Nandita Bose, *Walmart Goes to the Cloud to Close Gap with Amazon*, REUTERS (Feb. 14, 2018), https://www.reuters.com/article/us-walmart-cloud/walmart-goes-to-the-cloud-to-close-gap-with-amazon-idUSKCN1FY0K7 [https://perma.cc/K3AL-B3YU]; Farr & Levy, *supra* note 87.

[89] *See* Alex Hickey, *AWS Maintains Stranglehold on Cloud Market After Q2; Microsoft, Google Top Cloud Growth Rate*, CIO DIVE (July 30, 2018), https://www.ciodive.com/news/aws-maintains-stranglehold-on-cloud-market-after-q2-microsoft-google-top/528805/ [https://perma.cc/RH6N-VW6P]; Jamal Carnette, *Microsoft Is Taking on Amazon's Profit Center*, MOTLEY FOOL (Aug. 13, 2018), https://www.fool.com/investing/2018/08/13/microsoft-is-taking-on-amazons-profit-center.aspx [https://perma.cc/3MP9-9SK6].

too difficult to compete in such markets. This means customers could face fewer choices of SaaS and PaaS offerings.

Without any standards for interoperability, customers can find themselves locked into one of the dominant IaaS providers, making it difficult to switch if needed. Given the vibrant and dynamic state of the burgeoning SaaS market, this kind of market distortion could prevent new companies and technologies in downstream industries from reaching consumers. This is especially risky when the companies with market power in the IaaS market are global conglomerates that compete in a wide range of downstream industries.

V. INDUSTRY AND REGULATORY SOLUTIONS

A. U.S. ANTITRUST LAW FALLS SHORT IN ADDRESSING THE VERTICAL ANTICOMPETITIVE CONDUCT

United States antitrust laws fall short in addressing the vertical anticompetitive conduct discussed in this Note. When it comes to addressing the type of technology tying discussed in Section IV.B., U.S. courts have adopted a fact-specific approach, balancing the efficiencies and other benefits of tying with its anticompetitive effects.[90] Given the high bar *Twombly* pleading standards set for plaintiffs, and the difficulties plaintiffs face in accessing evidence of anticompetitive behavior, it is costly and timely for plaintiffs to bring antitrust cases.[91] In an article on the role of antitrust in broadband net neutrality, Hal Singer explained that "antitrust litigation imposes significant costs on private litigants, and it does not provide timely relief; if the net neutrality concern is a loss to edge innovation, a slow-placed [sic] antitrust court is not the right venue."[92]

The same goes for the possibility of lost innovation in cloud computing. Small to medium-sized competitors that are hurt by anticompetitive behavior in the IaaS market are not likely to pursue antitrust claims given the costly and time-consuming nature of antitrust litigation. If a new business is trying to enter the PaaS market with a novel service to compete against Amazon's Aurora database services, discriminatory treatment by Amazon in selling that start-up could mean that company never gets off the ground. Singer also emphasized, again in the context of broadband, that "competition is not the

[90] *See* United States v. Microsoft Corp., 253 F.3d 34, 59 (D.C. Cir. 2001) ("[C]ourts routinely apply a . . . balancing approach" requiring plaintiff to "demonstrate that the anticompetitive harm . . . outweighs the procompetitive benefit.").

[91] *See* Herbert J. Hovenkamp, *The Rule of Reason*, 70 FLA. L. REV. 81, 87–90 (2018).

[92] Hal J. Singer, *Paid Prioritization and Zero Rating: Why Antitrust Cannot Reach the Part of Net Neutrality Everyone Is Concerned About*, 17 ANTITRUST SOURCE 22, 23 (2017).

only value that net neutrality aims to address: end-to-end neutrality or nondiscrimination is a principle that many believe is worth protecting on its own."[93]

B. INDUSTRY-DRIVEN STANDARDIZATION IS CRITICAL TO ENSURING INTEROPERABILITY AND DATA PORTABILITY

Interoperability and data (and application) portability are critical to ensuring customers are not "locked-in" to a single IaaS provider or forced to purchase the IaaS provider's other cloud offerings. Without recognized standards, IaaS providers with market power can dictate which interfaces, and therefore which providers, can operate atop their cloud infrastructure. As recommended by NIST, U.S. government agencies should encourage the development and adoption of "voluntary consensus standards and in conformity assessment activities," to achieve interoperability and portability in cloud computing.[94] Just as standardization and federation concepts[95] enabled interoperability in the global telephone system and the Internet,[96] they too can help achieve cloud interoperability and data portability. Many international and domestic standards bodies, as well as industry consortia, are developing cloud interoperability standards.[97] For example, NIST and the Institute of Electrical and Electronic Engineers Standard Association (IEEE-SA) are partnering to address intercloud interoperability and create "an open, transparent infrastructure amongst cloud providers to support evolving technological and business models."[98] This effort aims to define "topology, functions, and

[93] *Id.*

[94] NAT'L INST. OF STANDARDS & TECH., NIST CLOUD COMPUTING STANDARDS ROADMAP 76 (2013) [hereinafter NIST CLOUD COMPUTING STANDARDS ROADMAP].

[95] *Federated Cloud*, NAT'L INST. OF STANDARDS & TECH., https://collaborate.nist.gov/twiki-cloud-computing/bin/view/CloudComputing/CloudFederated [https://perma.cc/D9MF-B5TK] (last visited Apr. 2, 2019) ("A Federation is multiple computing and/or network providers agreeing upon standards of operation in a collective fashion.").

[96] Press Release, IEEE Standards Ass'n, IEEE and National Institute of Standards and Technology (NIST) Team on Standards Development for Intercloud Interoperability and Federation (July 25, 2017), https://standards.ieee.org/news/2017/intercloud_interoperability_and_federation.html [https://perma.cc/NUP8-LVRW] [hereinafter IEEE Press Release] (IEEE and National Institute of Standards and Technology (NIST) Team on Standards Development for Intercloud Interoperability and Federation, Collaboration between NIST and IEEE P2302™ will help build consensus on creating an Intercloud—an open, transparent infrastructure amongst cloud providers to support evolving technological and business models).

[97] *See Industry Standards for Cloud*, CLOUD INDUSTRY F., https://www.cloudindustryforum.org/content/industry-standards-cloud [https://perma.cc/TWB6-MKL4] (last visited Apr. 2, 2019).

[98] IEEE Press Release, *supra* note 96.

governance for cloud-to-cloud interoperability and federation."[99] NIST recommends that U.S. government agencies encourage the adoption of such standards by actively participating in standards development, specifying such standards in the agencies' own procurements and grant guidance, and recommending specific cloud computing standards and best practices for government use.[100]

C. Interoperability Is Also Dependent on Service Level Agreements and Intellectual Property Law

Some of this platform standardization can also be achieved through the standardization of cloud Service Level Agreements, or "SLAs," which serve as a blueprint and a warranty for the scope and details of the cloud services to be provided.[101] These agreements usually detail the extent of the cloud customer's access to the data, the portability of the data, and the exit strategy should the customer want to transition to a different provider.[102] It is ultimately up to the customer, however, to be informed and vigilant when entering into SLAs. While sophisticated cloud customers may be increasingly savvy enough to demand important data portability provisions in their SLA's, the fewer IaaS providers in the market, the less leverage customers will have to negotiate SLAs that address their potential needs to switch providers.

Intellectual property law also has a role to play in making interoperability possible. A dominant cloud IaaS provider holding intellectual property rights to specific Application Programming Interfaces (APIs),[103] for example, might use those rights to restrict the compatible PaaS and SaaS applications that may operate atop of its infrastructure, limiting competition. The Federal Circuit examined the question of whether APIs are subject to copyright in *Oracle America, Inc. v. Google Inc.*, holding that "the declaring code and the structure, sequence, and organization of the API packages are entitled to copyright

[99] *Standard for Intercloud Interoperability and Federation (SIIF) Project Details*, IEEE STANDARDS ASSOCIATION, https://standards.ieee.org/project/2302.html [https://perma.cc/7QEZ-3HK7] (last visited Apr. 2, 2019).

[100] *See* NIST CLOUD COMPUTING STANDARDS ROADMAP, *supra* note 94, at 3–4.

[101] *See Service Level Agreements in the Cloud: Who Cares?*, WIRED (Dec. 2011), https://www.wired.com/insights/2011/12/service-level-agreements-in-the-cloud-who-cares/ [https://perma.cc/XC7H-98QP].

[102] *See id.*

[103] An API (Application Programming Interfaces) is a specification of possible interactions that allow programs to communicate with each other. *See* Jonathan Freeman, *What Is an API? Application Programming Interfaces Explained*, INFOWORLD (May 9, 2018), https://www.infoworld.com/article/3269878/apis/what-is-an-api-application-programming-interfaces-explained.html [https://perma.cc/9FYC-5PDV].

protection."[104] While the court left open the possibility that a competitive desire to achieve commercial interoperability may be relevant to a fair use analysis, it stated that it was not relevant given the facts of this case.[105] Many software companies use APIs from dominant cloud providers' cloud services to ensure compatibility between products.[106] Whether Amazon would pursue an infringement suit is unpredictable, but some argue that the decision in *Oracle America, Inc. v. Google Inc.* creates an incentive for copyright trolls to pursue litigation.[107]

D. U.S. Regulators Might Look to Public Utility Regulations to Ensure Open and Fair Access to IaaS

As cloud computing becomes an integral part of disseminating information over the Internet, there is a question of whether the neutrality of IaaS providers deserves a higher level of scrutiny on the part of regulators. Lina Khan discusses the challenges facing our current antitrust framework at length in her article, arguing that "the current framework in antitrust—specifically its equating competition with 'consumer welfare,' typically measured through short-term effects on price and output—fails to capture the architecture of market power in the twenty-first century marketplace."[108] One approach offered by Kahn to solve this problem is implementing public utility policies, such as nondiscrimination policies that prohibit platforms from privileging their own goods and discriminating among downstream industries. Another potential solution is imposing common carrier obligations that ensure open and fair access to an essential service.[109]

Given the trends toward rapid consolidation in the IaaS market, regulators might look to public utility solutions to prevent ineffective competition from affecting competition in other downstream markets. Observers have seen many of the problems and potential problems inherent to the market structure of the IaaS market before in other markets, particularly in the context of traditional public utilities such as telecommunications and electricity. Regulators may find it helpful to take lessons learned from past approaches and apply them where appropriate in the cloud computing context.

[104] Oracle Am., Inc. v. Google Inc., 750 F.3d 1339, 1348 (Fed. Cir. 2014).

[105] *See* Oracle Am., Inc. v. Google LLC, 886 F.3d 1179, n. 11 (Fed. Cir. 2018).

[106] Klint Finley, *The Case That Never Ends: Oracle Wins Latest Round vs. Google*, WIRED (Mar. 27, 2018), https://www.wired.com/story/the-case-that-never-ends-oracle-wins-latest-round-vs-google/ [https://perma.cc/TU8J-3DSW].

[107] *See id.* ("This creates a tremendous incentive for lawyers and copyright trolls to look for litigation[.]") (quoting Electronic Frontier Foundation legal director Corynne McSherry).

[108] Lina M. Kahn, *Amazon's Antitrust Paradox*, 126 YALE L.J. 710, 716 (2017).

[109] *See id.* at 799.

One example is regulatory intervention in the form of mandated access to ensure third-party providers—including managed service providers, PaaS providers, and SaaS providers—have non-discriminatory access to cloud infrastructure. This mandatory non-discriminatory access might parallel the local loop unbundling in the Telecommunications Act of 1996, which required incumbent local exchange carriers (ILECs) to open up their "last mile" to new entrants.[110] This is also an approach taken by many countries around the world to encourage competition in Internet services through the decoupling of Internet services from the last mile network.[111] This mandated access would require IaaS providers to provide non-discriminatory access to other smaller cloud providers, allowing them to resell the basic access to infrastructure along with their own differentiated services.

At the very least, regulatory solutions used in the past, such as the unbundling discussed in this Section, provide a concrete starting point from which to start building regulatory solutions for potential problems in the cloud computing market.

VI. CONCLUSION

This Note's purpose is to illuminate the market structure of the burgeoning cloud IaaS market and the potential challenges facing industry and regulators in ensuring this essential service facilitates competition and innovation in the varied markets that will come to depend on access to IaaS. This Note merely scratches the surface in the discussion of all the likely challenges—and possible solutions—facing regulators in the context of cloud IaaS. Other important challenges in the cloud IaaS market include predatory pricing, data privacy, cybersecurity, and platform reliability. The article aimed to highlight that the future impact of cloud computing on the economy, and society, depends largely on what is done now to safeguard the principles of openness and accessibility that enabled its creation. Cloud computing promises to continue to be transformative in many ways, but in the wake of exciting and disruptive change, it is important regulators remain steadfast in the commitment to ensuring tomorrow's innovators can reach consumers with tomorrow's life-changing innovation.

[110] *See* Emily Stewart, *Net Neutrality Isn't the Only Way to Keep the Internet Fair. It's Just the Only Way in America*, VOX (Dec. 14, 2017), https://www.vox.com/policy-and-politics/2017/12/14/16692318/net-neutrality-local-loop-broadband-internet-access [https://perma.cc/2HP5-PC2D].

[111] *See* Peter Bright, *We Don't Need Net Neutrality; We Need Competition*, ARS TECHNICA (June 2014), https://arstechnica.com/tech-policy/2014/06/we-dont-need-net-neutrality-we-need-competition/ [https://perma.cc/W5YN-HAWP].

RETHINKING EXPLAINABLE MACHINES: THE GDPR'S "RIGHT TO EXPLANATION" DEBATE AND THE RISE OF ALGORITHMIC AUDITS IN ENTERPRISE

Bryan Casey,[†] *Ashkon Farhangi*[††] *& Roland Vogl*[†††]

ABSTRACT

The public debate surrounding the General Data Protection Regulation's (GDPR) "right to explanation" has sparked a global conversation of profound social and economic significance. But from a practical perspective, the debate's participants have gotten ahead of themselves. In their search for a revolutionary new data protection within the provisions of a single chapter of the GDPR, many prominent contributors to the debate have lost sight of the most revolutionary change ushered in by the Regulation: the sweeping new enforcement powers given to European data protection authorities (DPAs) by Chapters 6 and 8 of the Regulation. Unlike the 1995 Data Protection Directive that it replaced, the GDPR's potent new investigatory, advisory, corrective, and punitive powers granted by Chapters 6 and 8 render DPAs de facto interpretive authorities of the Regulation's controversial "right to explanation." Now that the DPAs responsible for enforcing the right have officially weighed in, this Article argues that at least one matter of fierce public debate can be laid to rest. The GDPR provides a muscular "right to explanation" with sweeping legal implications for the design, prototyping, field testing, and deployment of automated data processing systems. The protections enshrined within the right may not mandate transparency in the form of a complete individualized explanation. But a holistic understanding of the interpretation by DPAs reveals that the right's true power derives from its synergistic effects when combined with the algorithmic auditing and "data protection by design" methodologies codified by the Regulation's subsequent chapters. Accordingly, this Article predicts that algorithmic auditing and "data protection by design" practices will likely become the new gold standard for enterprises deploying machine learning systems both inside and outside of the European Union.

DOI: https://doi.org/10.15779/Z38M32N986

[†] Bryan Casey is a Lecturer in Law at Stanford Law School, Legal Fellow at the Center for Automotive Research at Stanford, and an affiliate scholar at the Stanford Machine Learning Group and CodeX: The Stanford Center for Legal Informatics.

[††] Ashkon Farhangi is a Product Manager at Google and a fellow at CodeX: The Stanford Center for Legal Informatics.

[†††] Roland Vogl is the Executive Director of CodeX: The Stanford Center for Legal Informatics and the Stanford Program in Law, Science and Technology and a Lecturer in Law at Stanford Law School.

TABLE OF CONTENTS

I. INTRODUCTION

The year is 1995 and a spate of pioneering companies, including the upstarts Amazon.com and eBay, are staking their financial futures on an emerging technology that appears poised to forever transform the computing and communications worlds.[1] The technology, known among its proselytizers as the "Net," represents a new form of digital infrastructure that facilitates the worldwide sharing of data and communications without regard for geographic location.[2] Though adoption rates of this mysterious new technology remain relatively low, European anxieties surrounding its increasingly widespread use are already in full swing—precipitating the passage of legislation known as the Data Protection Directive (DPD) designed to grapple with the societal and technical complexities of a world on the cusp of a new digital era.[3]

Fast forward twenty years to the present and the Internet is, decidedly, old hat. But a technology equally alluring to the "Net" circa 1995 is enjoying a period of similarly rapid ascendance. The technology is known as "machine learning"[4]—or, for those of a more poetic bent, "artificial intelligence."[5] The level of optimism surrounding its potential to transform the world by turning

[1] *See* Harry McCracken, *1995: The Year Everything Changed*, FAST COMPANY (Dec. 30, 2015), https://www.fastcompany.com/3053055/1995-the-year-everything-changed [https://perma.cc/976P-XBMP]. eBay launched under the name of AuctionWeb at the time. *See id.*

[2] *See id.*

[3] *See* Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 1995 O.J. (L 281) 31 [hereinafter DPD]; Press Release, European Comm'n, IP/14/650, Commission Proposes a Comprehensive Reform of Data Protection Rules to Increase Users' Control of Their Data and to Cut Costs for Businesses (Jan. 25, 2012), http://europa.eu/rapid/press-release_IP-12-46_en.htm [https://perma.cc/T2X5-2526] [hereinafter GDPR Proposal].

[4] Machine learning can be described as a field of computer science that gives computers the ability to solve problems without being explicitly programmed to do so (i.e., the ability to "learn" by progressively improving performance on specific tasks). For references to definitions proffered by EU data authorities, see, e.g., DATATILSYNET, ARTIFICIAL INTELLIGENCE AND PRIVACY (Jan. 2018) [hereinafter ARTIFICIAL INTELLIGENCE AND PRIVACY]; NAT'L RESEARCH COUNCIL ET AL., FRONTIERS IN MASSIVE DATA ANALYSIS 101 (2013).

[5] While it is not wholly accurate to define "machine learning" and "artificial intelligence" as coextensive, for practical purposes this Article adopts to the convention of treating the two terms as synonymous. *See* ARTIFICIAL INTELLIGENCE AND PRIVACY, *supra* note 4, at 5 (defining artificial intelligence as "the concept used to describe computer systems that are able to learn from their own experiences and solve complex problems in different situations – abilities we previously thought were unique to mankind"). "Artificial intelligence is an umbrella term that embraces many different types of machine learning." *Id.* at 6.

machines into "intelligent"[6] decision-makers is matched only by the level of anxiety felt by those who fear the potential for bias to infiltrate machine decision-making systems once humans are removed from the equation.[7]

As recently as a decade ago, concerns surrounding bias within these types of complex automated systems would likely have struck many observers as far-fetched. Ever since the birth of computation with Alan Turing, humans have ascribed a kind of perfect "objectivity" to the mechanistic processes underlying algorithmic decision-making—a propensity now known as "automation bias."[8] Indeed, study after study has documented an innate human tendency to assume the validity of decisions made by algorithms,[9] even when presented with information that directly contradicts the decision's apparent validity.[10] The drafters of Europe's DPD explicitly acknowledged this phenomenon in 1992. They were so worried that "machine[s] using more and more sophisticated software" might be perceived as having "an apparently objective and incontrovertible character" that they felt it necessary to legislate specific

[6] The word intelligent, here, is used in quotes because of the fraught definitional issues associated with the term. As the scholar, Ryan Calo, notes, "Few complex technologies have a single, stable, uncontested definition [and] [r]obots are no exception." Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 529 (2015). For stylistic purposes, this Article uses "machine learning" and "artificial intelligence" interchangeably. Both terms lack a universally accepted definition, but this Article uses them to refers broadly to any "computerized system that exhibits behavior that is commonly thought of as requiring intelligence." EXEC. OFFICE OF THE PRESIDENT NAT'L SCI. & TECH. COUNCIL COMM. ON TECH., PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE 6 (2016).

[7] *See infra* Section IV.C and accompanying text.

[8] *See, e.g.*, A HISTORY OF ALGORITHMS: FROM THE PEBBLE TO THE MICROCHIP (Evelyn Barbin & Jean-Luc Chabert eds., 1999) [hereinafter A HISTORY OF ALGORITHMS]; Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1271–72 (2008); Kate Goddard et al., *Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators*, 19 J. AM. MED. INFORMATICS ASS'N 121 (2012); Christian Sandvig, *Seeing the Sort: The Aesthetic and Industrial Defence of "the Algorithm"*, 10 J. NEW MEDIA CAUCUS 1 (2014); Linda J. Skitka et al., *Accountability and Automation Bias*, 52 INT'L J. HUMAN COMPUTER STUD. 701, 704 (2000); Mary Cummings, *Automation Bias in Intelligent Time Critical Decision Support Systems*, AIAA 1ST INT. SYS. TECHNICAL CONF. (2004).

[9] An "algorithm" can be defined as "a formally specified sequence of logical operations that provides step-by-step instructions for computers to act on data and thus automate decisions." Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 674 n.10 (2016) (quoting SOLON BAROCAS ET AL., DATA & CIVIL RIGHTS: TECHNOLOGY PRIMER (2014)); *see* A HISTORY OF ALGORITHMS, *supra* note 8, at 2 (defining "algorithm" even more broadly as "any process of systematic calculation, that is a process that could be carried out automatically").

[10] *See* Cummings, *supra* note 8; Kathleen Mosier et al., *Automation Bias: Decision Making and Performance in High-Tech Cockpits*, 8 INT'L J. AVIATION PSYCHOL. 47, 47 (1997); Goddard et al., *supra* note 8, at 121.

measures guarding against it.[11]

In recent years, however, society's deferential attitude toward algorithmic objectivity has begun to wane—thanks, in no small part, to a flurry of influential publications examining bias within complex computational systems.[12] Particularly in the last five years, numerous studies across multiple industry sectors and social domains have revealed the potential for algorithmic systems to produce disparate real world impacts on vulnerable groups.[13] These

[11] *See Amended Proposal for a Council Directive on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data*, at 26, COM (1992) 422 final—SYN 297 (Oct. 15, 1992).

[12] *See, e.g.*, Bart Custers, *Data Dilemmas in the Information Society: Introduction and Overview*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 3, 20 (Bart Custers et al. eds., 2013); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 101 (2014) (noting "housing providers could design an algorithm to predict the [race, gender, or religion] of potential buyers or renters and advertise the properties only to those who [meet certain] profiles"); Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55, 57 (2013); Brent Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, 3 BIG DATA & SOC'Y 1, at 7–9 (2016); Latanya Sweeney, *Discrimination in Online Ad Delivery*, 11 ACM Queue 10, 12–13 (2013); Shoshana Zuboff, *Big Other: Surveillance Capitalism and the Prospects of an Information Civilization*, 30 J. INFO. TECH. 75 (2015); Solon Barocas, *Data Mining and the Discourse on Discrimination*, (2014) (unpublished manuscript), https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOnDiscrimination.pdf [https://perma.cc/LQ6R-FJZQ]; *see also, e.g.*, Citron, *supra* note 8, at 1254 ("Although programmers building automated systems may not intend to engage in rulemaking, they in fact do so . . . . The resulting distorted rules effectively constitute new policy that can affect large numbers of people."); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 4, 13–16 (2014) ("Because human beings program predictive algorithms, their biases and values are embedded into the software's instructions . . . ."); Devah Pager & Hana Shepherd, *The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets*, 34 ANN. REV. SOC. 181, 184 (2008); Michal S. Gal, *Algorithms as Illegal Agreements*, 34 BERKELEY TECH. L.J. 67 (2019); Julia Angwin et al., *Facebook (Still) Letting Housing Advertisers Exclude Users by Race*, PROPUBLICA (Nov. 21, 2017), https://www.propublica.org/article/facebook-advertising-discrimination-housing-racesex-national-origin [https://perma.cc/5B5W-WYEH]; Julia Angwin & Terry Parris Jr., *Facebook Lets Advertisers Exclude Users by Race*, PROPUBLICA (Oct. 28, 2016), https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race [https://perma.cc/4QDV-HC92].

[13] *See, e.g.*, Bryan Casey, *Title 2.0: Discrimination in a Data Driven Society*, 2019 J.L. & MOBILITY 36 (2019); Christine L. Borgman, *Open Data, Grey Data, and Stewardship: Universities at the Privacy Frontier*, 33 BERKELEY TECH. L.J. 365 (2018); Kevin Werbach, *Trust, but Verify: Why the Blockchain Needs the Law*, 33 BERKELEY TECH. L.J. 487 (2018); Kate Crawford, *The Hidden Biases in Big Data*, HARV. BUS. REV. (Apr. 1, 2013), https://hbr.org/2013/04/the-hidden-biases-in-big-data [https://perma.cc/E95C-TUQU]; Alistair Croll, *Big Data Is Our Generation's Civil Rights Issue, and We Don't Know It*, SOLVE FOR INTERESTING (July 31, 2012), http://solveforinteresting.com/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it [https://perma.cc/K77Z-PK3L]; Moritz Hardt, *How Big Data Is Unfair*, MEDIUM (Sept. 26, 2014), https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

revelations, in turn, have had a pronounced effect on scholars, policymakers, industry leaders, and society *writ large*—often serving as a rallying cry for greater efforts to promote fairness, accountability, and transparency in the design and deployment of highly automated systems.[14]

Yet, despite society's recent shift in attitude toward these types of algorithmic systems, the inexorable march of machine learning "eating the world" is only accelerating.[15] Across a diverse array of industries—from private social networks to public sector courtrooms[16]—organizations are adopting

[https://perma.cc/ZTZ4-8EG5]; Nadya Labi, *Misfortune Teller*, ATLANTIC (Jan./Feb. 2012), http://www.theatlantic.com/magazine/archive/2012/01/misfortune-teller/308846 [https://perma.cc/V3VV-84YU]; Anders Sandberg, *Asking the Right Questions: Big Data and Civil Rights*, PRAC. ETHICS (Aug. 16, 2012), http://blog.practicalethics.ox.ac.uk/2012/08/asking-the-right-questions-big-data-and-civil-rights [https://perma.cc/V86T-9S2P]; Tanzina Vega, *New Ways Marketers Are Manipulating Data to Influence You*, N.Y. TIMES: BITS (June 19, 2013), https://bits.blogs.nytimes.com/2013/06/19/new-ways-marketers-are-manipulating-data-to-influence-you/ [https://perma.cc/P89Y-2967].

[14] *See, e.g.*, DEP'T FOR DIGITAL, CULTURE, MEDIA & SPORT, DATA ETHICS FRAMEWORK (Aug. 30, 2018), https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework [https://perma.cc/FS48-CPA3]; HOUSE OF COMMONS, SCIENCE AND TECHNOLOGY COMMITTEE , ALGORITHMS IN DECISION-MAKING INQUIRY LAUNCHED , 2018, HC 351 (UK); SPECIAL EUROBAROMETER 431, DATA PROTECTION (June 2015); EUROPEAN DATA PROTECTION SUPERVISOR (EDPS), MEETING THE CHALLENGES OF BIG DATA: A CALL FOR TRANSPARENCY, USER CONTROL, DATA PROTECTION BY DESIGN AND ACCOUNTABILITY (2015); *Report with Recommendations to the Commission on Civil Law Rules on Robotics* 2015/2103(INL) (Jan. 27, 2017), http://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html [https://perma.cc/UE64-BJA5]; HOUSE OF COMMONS, SCIENCE AND TECHNOLOGY COMMITTEE, ROBOTICS AND ARTIFICIAL INTELLIGENCE, 2016, HC 145 (UK); INFORMATION COMMISSIONER'S OFFICE, BIG DATA, ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND DATA PROTECTION (2017) (UK); *see also* INFORMATION COMM'R'S OFFICE, OVERVIEW OF THE GENERAL DATA PROTECTION REGULATION (GDPR) (2017) (UK) [hereinafter ICO'S OVERVIEW OF GDPR]; NAT'L SCI. & TECH. COUNCIL, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE (2016); THE ROYAL SOCIETY, MACHINE LEARNING: THE POWER AND PROMISE OF COMPUTERS THAT LEARN BY EXAMPLE (2017); WETENSCHAPPELIJKE RAAD VOOR HET REGERINGSBELEID [DUTCH SCIENTIFIC COUNCIL FOR GOVERNMENT POLICY (WRR)], BIG DATA IN EEN VRIJE EN VEILIGE SAMENLEVING [BIG DATA IN A FREE AND SAFE SOCIETY], WRR-Rapport 95 (2016); Sophie Curtis, *Google Photos Labels Black People as 'Gorillas'*, TELEGRAPH (May 4, 2017), http://www.telegraph.co.uk/technology/google/11710136/Google-Photos-assigns-gorilla-tag-to-photos-of-black-people.html [https://perma.cc/25QY-TR9L].

[15] *See* Tom Simonite, *Nvidia CEO: Software Is Eating the World, but AI Is Going to Eat Software*, MIT TECH. REV. (May 12, 2017), https://www.technologyreview.com/s/607831/nvidia-ceo-software-is-eating-the-world-but-ai-is-going-to-eat-software/ [https://perma.cc/VT63-YSTL].

[16] *See, e.g.*, Corbett-Davies et al., Algorithmic Decision Making and the Cost of Fairness (June 10, 2017) (unpublished manuscript), https://arxiv.org/pdf/1701.08230.pdf [https://perma.cc/329E-WYRD]; Nikolaj Tollenaar et al., *StatRec —Performance, Validation and Preservability of a Static Risk Prediction Instrument*, 129 BULL. SOC. METHODOLOGY 25 (2016)

machine learning systems at unprecedented rates due to the technology's ability to radically improve data-driven decision-making at a cost and scale incomparable to that of humans.[17] Today, many agree that machine learning algorithms processing vast troves of data will only continue to play an increasingly large role in regulating our lives.[18] The question, thus, becomes: how are we to regulate these algorithms?

In 2016, the European Union sought to become a global pioneer in answering this question by replacing its 1990s-era DPD with comprehensive reform legislation known as the General Data Protection Regulation (GDPR).[19] The numerous protections introduced by the GDPR included an update to the DPD's rights surrounding automated decision-making.[20] The update formally enshrined what has since come to be referred to as the "right to explanation."[21] The right mandates that entities handling the personal data of EU citizens "ensure fair and transparent processing."[22] This requires providing citizens with access to "meaningful information about the logic involved" in certain automated decision-making systems.[23]

Many view the GDPR's "right to explanation" as a promising new mechanism for promoting fairness, accountability, and transparency in a world pervaded by complex algorithmic systems that can be difficult for observers to understand.[24] But as is true of numerous other rights enshrined within the GDPR, the precise contours of the "right to explanation" protections are less than clear—leading some commenters to wonder exactly how it will impact

(detailing published UK and Dutch predictive models involving recidivism).

[17] *See* Corbett-Davies et al., *supra* note 16.

[18] *See, e.g.*, Gideon Lewis-Kraus, *The Great A.I. Awakening*, N.Y. TIMES MAG. (Dec. 14, 2016), https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html [https://perma.cc/KG5C-NAD4].

[19] *See* Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 [hereinafter GDPR].

[20] *See id.*

[21] *See infra* Part II and accompanying notes.

[22] GDPR, *supra* note 19, at Recital 71.

[23] *Id.* at art. 15.

[24] *See infra* Section IV.C and accompanying notes; *see also, e.g.*, EXEC. OFF. OF THE PRESIDENT NAT'L SCI. & TECH. COUNCIL, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE (Oct. 2016); Catherine Stupp, *Commission to Open Probe into Tech Companies' Algorithms Next Year*, EURACTIV (Nov. 7, 2016), https://www.euractiv.com/section/digital/news/commission-to-open-probe-into-tech-companies-algorithms-next-year/ [https://perma.cc/B4TE-EHNQ]; GOV'T OFF. FOR SCI., ARTIFICIAL INTELLIGENCE: OPPORTUNITIES AND IMPLICATIONS FOR THE FUTURE OF DECISION MAKING (2016).

the use of machine learning in enterprise.[25]

In the two years since the GDPR's official publication, this uncertainty has ignited a heated global debate surrounding the Regulation's actual substantive protections.[26] The debate has centered on a cluster of four provisions found in Chapter 3 of the Regulation that circumscribe the specific text giving rise to the right. Scholars, industry leaders, and media sources across the globe have scoured the language of these provisions, proffering various competing interpretations of what the GDPR's new, and potentially revolutionary, "right to explanation" entails.[27] But lost in the debate's focus on the text of the provision has been a recognition of the more revolutionary change ushered in by the GDPR: the sweeping new enforcement powers given to Europe's data protection authorities.[28]

Unlike the DPD that it replaced, the GDPR grants EU data authorities vastly enhanced investigatory powers, a broad corrective "tool kit," and the capacity to levy fines several thousand times larger than the previous maximum limit.[29] Thanks to the GDPR's introduction of these truly threatening administrative powers, EU data authorities will no longer be rendered the toothless watchdogs many companies have long viewed them to be.[30] Rather, these newly empowered authorities will play a weighty role in enforcing and,

[25] *See infra* Part II.B and accompanying notes.

[26] *See infra* Part II and accompanying notes.

[27] *See infra* Part III and accompanying notes; *see also, e.g.*, FRANCESCA ROSSI, ARTIFICIAL INTELLIGENCE: POTENTIAL BENEFITS AND ETHICAL CONSIDERATIONS (2016). For media perspectives, see Cade Metz, *Artificial Intelligence Is Setting Up the Internet for a Huge Clash With Europe*, WIRED (July 11, 2016), https://www.wired.com/2016/07/artificial-intelligence-setting-internet-huge-clash-europe/ [https://perma.cc/4JSZ-THTR]; Bernard Marr, *New Report: Revealing The Secrets of AI or Killing Machine Learning?*, FORBES (Jan. 12, 2017), https://www.forbes.com/sites/bernardmarr/2017/01/12/new-report-revealing-the-secrets-of-ai-or-killing-machine-learning/#35a503e543ef [https://perma.cc/K8UQ-Q3GA]; Liisa Jaakonsaari, *Who Sets the Agenda on Algorithmic Accountability?*, EURACTIV (Oct. 29, 2016), https://www.euractiv.com/section/digital/opinion/who-sets-the-agenda-on-algorithmic-accountability/ [https://perma.cc/938H-4TPR]; Nick Wallace, *EU's Right to Explanation: A Harmful Restriction on Artificial Intelligence*, TECHZONE360 (Jan. 25, 2017), http://www.techzone360.com/topics/techzone/articles/2017/01/25/429101-eus-right-explanation-harmful-restriction-artificial-intelligence.htm [https://perma.cc/7XEA-B834].

[28] *See* GDPR, *supra* note 19, at chs. 6, 8.

[29] *See id.* The exact multiple can vary depending on the company's annual turnover. *See infra* Part III.

[30] *See* Natasha Lomas, *WTF Is GDPR*, TECHCRUNCH (Jan. 2018), https://techcrunch.com/2018/01/20/wtf-is-gdpr/ [https://perma.cc/G9FD-LRQV] (noting that the "beefing up of enforcement that's baked into the new regime means there's a better opportunity for DPAs to start to bark and bite like proper watchdogs"); *infra* Part III and accompanying notes.

therefore, *interpreting* the GDPR's numerous protective mandates.[31]

Viewed through this lens, it becomes apparent that many disagreements surrounding the "right to explanation" may have clearer answers than the current state of debate suggests. While vocal observers on both sides have dominated the headlines, those tasked with actually enforcing the "right to explanation" have quietly gone to work.[32] In the last six months, these authorities have produced a richly detailed framework for companies seeking to promote compliance with the GDPR's "right to explanation."[33] Given that these are the very same authorities on the front lines of enforcing compliance, their interpretations merit careful consideration.

Now that the dust from this recent burst of activity by data authorities has begun to settle, this Article attempts to take stock of the new developments—just in time for the Regulation's recent effectuation. In doing so, this Article seeks to turn the page within the GDPR's fraught "right to explanation" debate by answering a question that has, thus far, gone almost entirely overlooked: What do those actually tasked with enforcing the right think it entails?

Stepping outside of the debate's focus on the text of the GDPR, this Article adopts a holistic approach to understanding the Regulation's somewhat loosely-worded mandate. This Article contextualizes the "right to explanation" provisions by setting them against the backdrop of the potent range of new administrative capabilities prescribed by subsequent provisions. These new provisions effectively render Europe's data protection agencies de facto interpretive authorities.[34] In adopting this approach, this Article takes particular pains to let the words of the Regulation and its downstream interpreters speak for themselves—making use of direct quotes or passages whenever possible.[35]

Through the words of the authorities in charge of enforcing the GDPR, this Article finds a muscular "right to explanation" enshrined within the Regulation—albeit one that is subtly different from the competing visions contemplated by some scholars and industry experts. Europe's data protection authorities consistently reveal that they envisage the "right to explanation" not only as an individual remedial mechanism but also as part and parcel of a broader form of oversight with broad implications for the design and

[31] *See id.*

[32] *See infra* Part IV.

[33] *See infra* Part IV and accompanying notes.

[34] *See infra* Part III and accompanying notes.

[35] The hope, here, is to minimize editorializing—not to bore the reader with block quotes.

deployment of automated systems that process personal data.[36]

This Article seeks to better understand this newly articulated "right to explanation" and, in doing so, hopes to shed light on how enterprises can prepare for, react to, and promote compliance with what will doubtless be one of the most influential data protection frameworks of the coming decades. The Article proceeds in five parts. Part II traces the history of the public debate surrounding the "right to explanation." It begins with the right's origins in the specific text of Chapter 3 and proceeds to overview several of the most prominent contributions to the public debate thus far. In highlighting the debate's merits and demerits, it argues that the participants' general failure to countenance the substantive changes to enforcement introduced by Chapters 6 and 8 of the Regulation represents a fundamental oversight—one that has hindered a genuine understanding of the right's substantive protections.

Part III turns the page in the debate by broadening its focus to include Chapters 6 and 8 of the GDPR. It contextualizes the newfound role that enforcement agencies will play by detailing their limitations under the DPD and outlining their vastly enhanced administrative powers granted by Chapters 6 and 8. It argues that these newly empowered data watchdogs will serve as *functional* interpretive authorities of the GDPR's "right to explanation," even if other legislative or judicial authorities may, theoretically, have the final say. Because these agencies will be on the front lines of enforcement, their interpretations will, of necessity, be the most relevant for enterprises seeking to comply with the GDPR. Fortunately, these very agencies have recently produced extensive guidance describing their interpretations of the "right to explanation" that offers powerful insights into the substantive protections afforded by the GDPR's vaguely-worded mandate.

Part IV details this newly issued guidance and summarizes its implications for companies seeking to better understand what compliance with the GDPR's "right to explanation" actually entails. It reveals that Europe's data authorities have repeatedly envisioned the "right to explanation" as a robust data protection whose true power lies in its synergistic combination with the "data protection by design" principles codified in the Regulation's subsequent chapters. As a result, this Article argues that data auditing methodologies designed to safeguard against algorithmic bias throughout the entire product life cycle will likely become the new norm for promoting compliance in automated systems. It further argues that this more general version of a "right to explanation" offers greater hope of promoting genuine "algorithmic accountability" than the individualized remedial mechanism many

[36] *See infra* Part III and accompanying notes.

commentators have presumed it to be.

Part V examines the GDPR's global implications for companies and countries grappling with compliance, both inside and outside of Europe. It argues that the Regulation will likely have an outsized extraterritorial impact due to the well-documented "Brussels Effect" and the introduction of several legal mechanisms that implicate entities operating outside of the EU. Thanks to the far-flung legal reach of the Regulation, it argues that the "right to explanation"—as envisioned by the GDPR's enforcement authorities—appears destined to become part of a new global data protection standard for companies handling personal information. The new standard will certainly pose its share of challenges for enterprises seeking to deploy sophisticated algorithms. But it also offers those who hope for a more fair, accountable, and transparent automated decision-making systems genuine reason for optimism.

II. DOES THE GDPR ENVISAGE A RIGHT TO EXPLANATION?

In January 2012, the European Commission made global headlines by submitting a proposal to "update and modernise the principles enshrined in the 1995 Data Protection Directive."[37] For seventeen years, the DPD had reigned as Europe's preeminent legislation governing the processing of digital data. But after nearly two decades, the longstanding Directive was beginning to show signs of age. The DPD was originally passed when "less than 1% of Europeans used the internet."[38] Since then, the Commission noted, "[t]echnological progress . . . [had] profoundly changed the way [] data is collected, accessed and used."[39]

The press release accompanying the Commission's announcement set the stage for "a comprehensive reform of [the DPD's] data protection rules."[40] The Commission called for rules to be designed "to increase users' control of their data," to "provide[] for increased responsibility and accountability for those processing personal data," and to create a "single set of rules" that would be "valid across the EU."[41] More than three years of negotiations followed the preliminary proposal, eventually culminating in the formal adoption of the General Data Protections Regulation (GDPR) in April of 2016.[42] The finalized

[37] GDPR Proposal, *supra* note 3.

[38] *Id.*

[39] *Id.*

[40] *Id.*

[41] *Id.*; CONSOLIDATED VERSION OF THE TREATY ON THE FUNCTIONING OF THE EUROPEAN UNION art. 288, 2008 O.J. C 115/47.

[42] *See* GDPR, *supra* note 19.

Regulation constituted a major overhaul of European data processing standards. By enumerating a litany of powerful protections, the new Regulation intended to make the EU bloc "fit for the digital age."[43]

One such protection—located within Chapter 3 of the GDPR—sets forth what the Regulation describes as the "right not to be subject to a decision based solely on automated processing."[44] The protection establishes a number of safeguards designed to ensure the "fair and transparent processing" of personal data, including an obligation that entities provide "meaningful information about the logic involved" in certain types of highly automated decision-making systems.[45] The protection's requirement that "meaningful information" be made available to data subjects has led it to be variously characterized as enshrining a "right to information," a "right to be informed," or, most commonly, a "right to explanation."[46]

As the first piece of European Union Regulation to explicitly gesture toward such a right,[47] the substantive protections that eventually flow from it will set a precedent with ramifications extending far beyond the technology sector. While the usual suspects, such as Facebook, may have grabbed global headlines by announcing millions of dollars spent toward promoting GDPR compliance, the rapid proliferation of machine learning technology across diverse industries indicates that vast swaths of the private sector will soon be forced to take action. Depending on how the protection is eventually applied

[43] EUROPEAN COMM'N, REFORM OF EU DATA PROTECTION RULES (2018), http://ec.europa.eu/justice/data-protection/reform/index_en.htm [https://perma.cc/JH2B-YRMU].

[44] GDPR, *supra* note 19, at art. 22; *see id.* at arts. 13(2)(f), 14(2)(g), 15(1)(h). As is likely obvious, this phrasing leaves open considerable room for ambiguity.

[45] *See id.* at arts. 14(2), 14(2)(g).

[46] *See, e.g.*, Bryce Goodman & Seth Flaxman, EU Regulations on Algorithmic Decision Making and "a Right to an Explanation" (June 28, 2016) (unpublished manuscript), https://ora.ox.ac.uk/objects/uuid:593169ee-0457-4051-9337-e007064cf67c/download_file?safe_filename=euregs.pdf&file_format=application%2Fpdf&type_of_work=Journal+article [https://perma.cc/C6UP-DZQE]; Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76 (2017); Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT'L DATA PRIVACY L. 233 (2017); *Data Subjects' Rights*, RADBOUD U., https://www.ru.nl/privacy/english/protection-personal-data/data-subjects-rights/#hf4dfc431-41bd-452c-8cac-3f98083db3b1 [https://perma.cc/8KQ2-WUFM]; ARTICLE 29 WORKING PARTY, GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING AND PROFILING FOR THE PURPOSES OF REGULATION 2016/679, 9 (2017) [hereinafter A29WP Automated Decision-Making Guidelines].

[47] This could be more precisely phrased as the European Union Regulation to mandate this right in the context of *automated systems with a meaningful threat of enforcement*—a nuance that is covered in greater detail in Part III *infra*.

in practice, it could have profound implications for the use of some of the most powerful computational techniques available to modern enterprises. But, as is true of many protections enshrined within the legislative text of the GDPR, the precise reach of the right is far from certain. A careful examination of the language provides a useful starting point for understanding and contextualizing it.

A. SPECIFIC TEXT GIVING RISE TO THE "RIGHT TO EXPLANATION"

Article 22 of the GDPR grants all data subjects[48] a rebuttable[49] "right not to be subject to a decision based solely[50] on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."[51] The GDPR defines "processing" as follows:

> [A]ny operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction[.][52]

The GDPR's use of the term "profiling" introduces a relatively novel concept under EU data protection law.[53] The regulation defines "profiling" as

> any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements[.][54]

[48] *See* GDPR, *supra* note 19, at art. 4. The GDPR defines a "data subject" as "an identified or identifiable natural person" and "an identifiable natural person" as "one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person." *Id.*

[49] *See id.* at art. 22(2)–(4) (specifying limited circumstances where automated decision-making is permitted, and providing for different data safeguards).

[50] This term has recently been subject to clarification. *See infra* Part IV.

[51] GDPR, *supra* note 19, at art. 22(1).

[52] *Id.* at art. 4(2).

[53] *See* Frederike Kaltheuner & Elettra Bietti, *Data is Power: Towards Additional Guidance on Profiling and Automated Decision-Making in the GDPR*, 2 J. INFO. RIGHTS, POL'Y & PRACTICE (2018).

[54] GDPR, *supra* note 19, at art. 4(4). Recital 71 of the GDPR adds:

> Such processing includes 'profiling' that consists of any form of automated

Article 22(2) enumerates a limited number of circumstances in which companies[55] processing personal data are exempt from its prohibitions—including when automated decision-making is done consensually or is necessary for contracting.[56] But even in such instances, Article 22 requires that companies nevertheless "implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests."[57] This requirement, at a minimum, includes the subject's "right to obtain human intervention on the part of the [company], to express his or her point of view and to contest the decision."[58]

Article 22's protections are buttressed by those located within Articles 13–15, pertaining to the rights of data subjects whose personal information is directly or indirectly implicated by automated processing techniques. These Articles are intended to "provide the data subject with the . . . information necessary to ensure fair and transparent processing."[59] In fulfilling this goal, Articles 13(2)(f), 14(2)(g), and 15(1)(h) mandate that companies provide subjects with information regarding "the existence of automated decision-making, including profiling, referred to in Article 22 . . . and, at least in those cases, *meaningful information about the logic involved*, as well as the significance and the envisaged consequences of such processing for the data subject."[60]

In addition to the text of the GDPR, the accompanying nonbinding Recital

> processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her.

See GDPR, *supra* note 19, at Recital 71; *see also* Mireille Hildebrandt, *Defining Profiling: A New Type of Knowledge?*, *in* PROFILING THE EUROPEAN CITIZEN 17 (Mireille Hildebrandt & Serge Gutwirth eds., Springer 2008) (exploring the difference between organic and machine profiling).

[55] *See* GDPR, *supra* note 19, at art. 4(7). The GDPR does not single out companies, but instead uses the term "controller" which "means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law." *Id.*

[56] *See id.*, at art. 22(2)–(4).

[57] *Id.* at art. 22.

[58] *Id.*

[59] *Id.* at arts. 13, 14.

[60] *Id.* at arts. 13(2)(f), 14(2)(g), 15(1)(h) (emphasis added). This disclosure requirement extends even to data subjects whose personal information has not been directly obtained by a company.

71 offers further clarification regarding the Regulation's protections pertaining to automated decision-making.[61] The Recital states that the data processing techniques implicating personal data "should be subject to suitable safeguards, which should include [the provision of] specific information to the data subject[,]" as well as the rights "to obtain human intervention," "to express his or her point of view," "to *obtain an explanation of the decision reached* after such assessment," and "to challenge the decision."[62] The Recital further stipulates:

> In order to ensure fair and transparent processing . . . [companies] should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject, and prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or processing that results in measures having such an effect.[63]

While the authority of the Recital is nonbinding under EU law, it nonetheless provides a critical reference point for future interpretations by data protection agencies as well as for co-determinations of positive law that

[61] *See* Tadas Klimas & Jurate Vaiciukaite, *The Law of Recitals in European Community Legislation*, 15 ILSA J. INT'L & COMP. L. 61, 62, 92 (2008). Recitals in EU law lack "independent legal value, but they can expand an ambiguous provision's scope. They cannot, however, restrict an unambiguous provision's scope, but they can be used to determine the nature of a provision, and this can have a restrictive effect." *Id.* at 63. "Recitals explain the background to the legislation and the aims and objectives of the legislation. They are, therefore, important to an understanding of the legislation which follows." COMMISSION OF THE EUROPEAN COMMUNITIES, GUIDE TO THE APPROXIMATION OF EUROPEAN UNION ENVIRONMENTAL LEGISLATION 115 (2017); *see* Case C-355/95 P, Textilwerke Deggendorf GmbH v. Comm'n, 1997 E.C.R. I-02549 ("In that regard, it should be stated that the operative part of an act is indissociably linked to the statement of reasons for it, so that, when it has to be interpreted, account must be taken of the reasons which led to its adoption."). European Court of Justice (ECJ) jurisprudence reveals that the role of Recitals is "to dissolve ambiguity in the operative text of a framework." Wachter et al., *supra* note 46, at 80. According to the ECJ: "Whilst a recital in the preamble to a regulation may cast light on the interpretation to be given to a legal rule, it cannot in itself constitute such a rule." Case 215/88, Casa Fleischhandels-GmbH v. Bundesanstalt fur Landwirtschaftliche Marktordnung, 1989 E.C.R 02789; *see* Roberto Baratta, *Complexity of EU Law in the Domestic Implementing Process*, *in* 2 THE THEORY AND PRACTICE OF LEGISLATION 293 (2014) (highlighting how the complexity of EU law can cause difficulties at the national level); Klimas & Vaiciukaite, at 62.

[62] GDPR, *supra* note 19, at Recital 71 (emphasis added).

[63] *Id.*

may be made by legislators, courts, or other authorities.[64]

B. THE "RIGHT TO EXPLANATION" DEBATE

Despite the GDPR's concerted efforts to detail the protections enshrined under Articles 13, 14, 15, and 22, much uncertainty continues to shroud the Regulation's so-called "right to explanation." This phenomenon owes, in large part, to the GDPR's somewhat fuzzy mandate that entities "ensure fair and transparent processing" by providing "meaningful information about the logic involved" in automated decision-making systems. At a minimum, the protection appears to envisage a limited right for data subjects to understand and verify the basic functionality of certain automated decision-making systems. But beyond that minimum threshold, the precise contours of the "right to explanation" have been the subject of much speculation—giving rise to an "explosive" public debate.[65]

Among the most prominent contributions to the debate, thus far, have been three distinct perspectives originating from scholars within the U.K. and the U.S.[66] Their claims and critiques are set forth below.

1. *The Original Claim*

Goodman's and Flaxman's conference paper—*European Union Regulations on Algorithmic Decision-making and a "Right to Explanation"*—first popularized the knotty, sometimes vexing, issues at the heart of the GDPR's "right to explanation."[67] Published just two months after the Regulation's official release, the piece drew widespread attention to the technical and societal challenges inherent in "explain[ing] an algorithm's decision" made by machine learning algorithms.[68] Goodman and Flaxman observed that, unlike algorithms

[64] These authorities, amongst others, include the GDPR's designated "Supervisory Authorities," the Article 29 Working Party, the European Data Protection Board, the European Data Protection Supervisor, and the European Data Protection Supervisor's Ethics Advisory Group.

[65] *See infra* Section II.B.

[66] Many other contributors beyond these three have also thrown their hats in the ring.

[67] *See* Goodman & Flaxman, *supra* note 46. It should be noted that this paper was subsequently revised.

[68] *See* FRANK PASQUALE, THE BLACK BOX SOCIETY 3–4 (2015); Brenda Reddix-Smalls, *Credit Scoring and Trade Secrecy: An Algorithmic Quagmire or How the Lack of Transparency in Complex Financial Models Scuttled the Finance Market*, 12 U.C. DAVIS BUS. L. J. 87 (2011); Frank Pasquale, *Restoring Transparency to Automated Authority*, 9 J. ON TELECOMM. & HIGH TECH. L. 235, 237 (2011); Maayan Perel & Niva Elkin-Koren, *Accountability in Algorithmic Copyright Enforcement*, 19 STAN. TECH. L. REV. 473, 482 (2016); *see generally* NICHOLAS DIAKOPOULOS, ALGORITHMIC ACCOUNTABILITY REPORTING: ON THE INVESTIGATION OF BLACK BOXES (Tow Centre for Digital Journalism, 2013); Goodman & Flaxman, *supra* note 46.

of past decades,[69] machine learning systems in increasingly widespread usage were "alone on the spectrum in their lack of interpretability."[70] The scholars noted an inherent "tradeoff between the representational capacity of a model and its interpretability"—one that sometimes rendered the underlying decision-making process of the most powerful systems an uninterpretable "black box."[71]

While these types of "black box" algorithms had existed in research labs since the 1980s, Goodman and Flaxman made the prescient observations that their recent proliferation throughout industry presented many challenges for companies and governments seeking to comply with the GDPR.[72]The scholars discussed how numerous factors—including potentially biased training sets, uneven "data quality," the complexity of the most powerful predictive models, and the steep barriers to technical fluency—could pose significant challenges for modern enterprises seeking to comply with the GDPR's mandate of algorithmic explicability.[73]

Although the scholars' work was widely credited with sparking the "right to explanation" debate,[74] their piece was less a legal treatise than a technical primer. Their analysis offered relatively little commentary regarding the right's substantive protections and made only a passing reference to the GDPR's

[69] I.e., those which relied on explicit, rules-based logic for processing information.

[70] *See* Goodman & Flaxman, *supra* note 46, at 6 (quoting PAULO J. G. LISBOA, INTERPRETABILITY IN MACHINE LEARNING PRINCIPLES AND PRACTICE 1521 (2013)).

[71] *See id.* Machine learning techniques that explicitly encode logic do exist—particularly in the natural language processing and bioinformatics realms—but are not focused on for purposes of concision.

[72] *See* Robert D. Hof, *Deep Learning*, MIT TECH. REV. (2013), https://www.technologyreview.com/s/513696/deep-learning [https://perma.cc/Y822-QJC9] (noting that in the mid-80s, "[scientists] spark[ed] a revival of interest in neural networks with so-called "deep" models that made better use of many layers of software neurons"); Goodman & Flaxman, *supra* note 46.

[73] "Data quality" is a broadly construed term whose components include "accuracy, precision, completeness, consistency, validity, and timeliness, though this catalog of features is far from settled." *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 684 n.47 (2016); Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 21 (2017); *see also, e.g.*, Luciano Floridi, *Information Quality*, 26 PHIL. & TECH. 1 (2013); Richard Y. Wang & Diane M. Strong, *Beyond Accuracy: What Data Quality Means to Data Consumers*, 12 J. MGMT. INFO. SYS. 5 (1996); LARRY P. ENGLISH, INFORMATION QUALITY APPLIED (2009).

[74] *See* Michelle Menting, *EU GDPR: The Impact on the Use of Machine Learning*, ABI RES. (Sept. 17, 2018), https://www.abiresearch.com/blogs/eu-gdpr-impact-use-machine-learning/ [https://perma.cc/8SXQ-V9V8] (crediting Goodman and Flaxman with initiating the debate); Selbst & Powles, *supra* note 46, at 234 (noting that the most "most prominent contributions" to the debate are Goodman and Flaxman's piece and Wachter et al.'s response).

newly introduced enforcement provisions. When the piece did discuss the "right to explanation" directly, Goodman and Flaxman construed the protection as relatively narrow. Aside from a single loosely-worded sentence in the paper's abstract that received outsized attention, the scholars suggested that the "right to explanation" could be satisfied relatively easily. They indicated that simply answering questions such as: "*Is the model more or less likely to recommend a loan if the applicant is a minority?*" or "*Which features play the largest role in prediction?*" could suffice.[75]

2. *The Response*

In response to the widespread attention garnered by Goodman's and Flaxman's conference paper, Wachter et al. entered into the public arena with the provocatively titled piece, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation.*[76] The scholars wasted no time going on the offensive, immediately calling into doubt both the legal existence and the technical feasibility of what Goodman and Flaxman referred to as the GDPR's "right to explanation." Wachter et al.'s contribution offered a richly detailed tour of the Regulation's relevant text and associated Recital—one that reached greater analytic depths than the technically-oriented conference paper it criticized. The scholars articulated a powerful framework for distinguishing questions of algorithmic explicability along chronological and functional dimensions—an important contribution that has since been replicated by numerous researchers.[77]

But as thorough as Wachter et al.'s analysis may have been, their focus was also highly selective. Several of their arguments all but ignored key terms within Articles 13, 14, 15, and 22. In particular, Wachter et al. disregarded the word "meaningful" as applied to a substantive analysis of the phrase "meaningful information about the logic involved" in automated decision-making.[78] Just as importantly, their piece paid short shrift to the Regulation's powerful new administrative capabilities. Instead, their discussion of the GDPR's new

[75] Goodman & Flaxman, *supra* note 46 (emphasis added). The scholars offered virtually no substantive support for their argument that the right could be satisfied with these types of explanations.

[76] Wachter et al., *supra* note 46.

[77] *See, e.g.*, Edwards & Veale, *supra* note 73. Wachter et al.'s framework distinguishes between explanations describing "system functionality" and "specific decisions," and also distinguishes between explanations that occur *before* a data-subject's information has been processed and those that occur *after*. *See* Wachter et al., *supra* note 46, at 78–79.

[78] *See* Wachter et al., *supra* note 46, at 84. The scholars also made a few claims of astonishing scope, including one assertion that, "There are no ambiguities in the language [of the GDPR] that would require further interpretation with regard to the minimum requirements that must be met by data controllers." *Id.* at 80.

enforcement capabilities was limited to a single footnote.[79] Most strikingly of all, the central thesis they advanced was outright contradicted by their own subsequent analysis. After electing to title their work *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*,[80] the scholars went on to repeatedly acknowledge that just such a right existed—noting, for example, that the Regulation could mandate "an explanation when automated decisions have (i) legal or similarly significant effects, and (ii) are based solely on automated processes."[81]

Rather than calling it a "right to explanation," however, the scholars instead sought to replace it with a phrase of narrower implications. They insisted that "the GDPR does not . . . implement a right to explanation, but rather [a] 'right to be informed.' "[82] The scholars, however, went on to note that this mandate provided data-subjects, at minimum, "a right to explanation of system functionality . . . [subject to] restrict[ions] by the interests of data controllers and future interpretations."[83] As such, their insistence on calling it a "right to be informed" appeared to be a distinction of little more than semantic significance.[84]

3. *The Rebuttal*

In November 2017—with the GDPR just six months away and the "right to explanation" debate rapidly rising to a fevered pitch—Selbst and Powles entered into the fray with a point-by-point takedown of Wachter et al. in *Meaningful Information and the Right to Explanation*. Their contribution sought to address what they described as the numerous "unfounded assumptions and unsettling implications of [Wachter et al.'s] analytical frame."[85] In doing so, Selbst and Powles "offer[ed] a positive conception of the right [to explanation] located in the text and purpose of the GDPR."[86] They convincingly argued that it "should be interpreted functionally, flexibly, and should, at a minimum, enable a data subject to exercise his or her rights under the GDPR and human

[79] *See id.* at 99 n.130.

[80] *See id.* The scholars Selbst and Powles correctly noted that this tactic was "not only disingenuous but dangerous, as it invites less scrupulous or more time-pressed advocates to cite the paper for the proposition that there is no right to explanation, which is not even what the paper argues in substance." Selbst & Powles, *supra* note 46, at 238.

[81] Wachter et al., *supra* note 46, at 95.

[82] *Id.* at 77.

[83] *Id.* at 96.

[84] *See* Selbst & Powles, *supra* note 46, at 239.

[85] *See id.*; *infra* Section II.B.2. Many of these criticisms are outlined in the section above.

[86] Selbst & Powles, *supra* note 46, at 234.

rights law."[87]

Selbst's and Powles's piece represented a vital course correction in a public debate that had begun to more closely resemble a rebranding effort than an actual refutation of the substantive right itself.[88] But their contribution occurred in advance of Europe's most influential data protection authorities releasing extensive guidance which provided much needed clarity on the hotly contested topic.[89] Accordingly, the actual language of EU data protection authorities that emerged immediately after its publication did not ground Selbst and Powles's piece. Further, the piece did little to underscore the GDPR's newly-invigorated enforcement measures, as well as the practical implications that flow from them, which are discussed below.

C. LOST IN THE FOG OF BATTLE

Since its origins with Goodman and Flaxman, the GDPR's "right to explanation" debate has fostered a conversation of profound global significance—exploring the economic benefits, technical feasibility, and social tradeoffs of applying "algorithmic accountability" practices in enterprise and government.[90] The contributions of Goodman, Flaxman, Selbst, Powles, and Wachter et al. constitute just a tiny sample of the vast and impressively diverse array of perspectives on this issue.[91] Over a period of just eighteen months, countless industry leaders, media sources, and researchers of various backgrounds have also contributed their unique perspectives.[92] But

[87] *Id.* at 242.

[88] Watcher et al.'s piece continues to enjoy widespread popularity among more casual observers—with many remaining unaware of the important counterweight provided by Selbst & Powles.

[89] *See* Selbst & Powles, *supra* note 46.

[90] *See infra* Section IV.C for a more detailed description of the literature on "algorithmic accountability."

[91] Mendoza and Bygrave, who argue that the "right to explanation" arises as a necessary precondition to Article 22(3)'s "right to contest" could also be added to this list, but are not discussed in detail for purposes of concision. Izak Mendoza & Lee A. Bygrave, *The Right Not to Be Subject to Automated Decisions Based on Profiling*, *in* EU INTERNET LAW: REGULATION AND ENFORCEMENT 77 (T.-E. Synodinou et al. eds., Springer 2017).

[92] *See, e.g.*, Rich Caruana et al., *Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission*, *in* KDD '15PROCEEDINGS OF THE 21ST ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1721 (2015); David Bamman, Interpretability in Human-Centered Data Science (2016) (unpublished manuscript), https://cscw2016hcds.files.wordpress.com/2015/10/bamman_hcds.pdf [https://perma.cc/3KLR-8MDY]; Michael Gleicher, *A Framework for Considering Comprehensibility in Modeling*, 4 BIG DATA 75 (2016); Finale Doshi-Valez & Been Kim, A Roadmap for a Rigorous Science of Interpretability (2017) (unpublished manuscript), https://arxiv.org/abs/1702.08608 [https://perma.cc/Q9K4-PZYM]; Eric Horvitz, Presentation at the Berkeley Center for Law & Technology: On the Meaningful Understanding

mystifyingly, many of the most distinguished contributions to this multifaceted debate have largely overlooked what is potentially the most profound change of all heralded by the GDPR: the sweeping new enforcement powers granted to EU data protection authorities by the new Regulation.

Beyond the "right to explanation" debate's narrow focus on Articles 13, 14, 15, and 22, there lies a series of provisions that appear destined to forever change the practical reality of enforcement by data protection authorities. These Articles—contained in Chapters 6 and 8 of the Regulation—grant vast new administrative powers to EU watchdog agencies that have long been viewed as toothless under the DPD.[93] Failing to elucidate the profound new role that these freshly empowered agencies will play in enforcing and, therefore, *interpreting* the GDPR's "right to explanation" currently represents a major blind-spot within the public debate. If left unaddressed, this blind spot risks allowing the public debate to move in an unproductive and unnecessarily adversarial direction.

III. TURNING THE PAGE IN THE "RIGHT TO EXPLANATION" DEBATE

Although the introduction of the GDPR will represent the largest overhaul of EU data protection laws in twenty years, the Regulation's most revolutionary change actually involves the addition of a host of new legal mechanisms for promoting enforcement.[94] After all, the EU has long boasted an extensive list of rules[95] that set a high bar for data protection, including

of the Logic of Automated Decision Making (Mar. 24 2017); Ethan Chiel, *EU Citizens Might Get a 'Right to Explanation' About the Decisions Algorithms Make*, SPLINTER (July 5, 2016), http://fusion.kinja.com/eu-citizens-might-get-aright-to-explanation-about-the-1793859992 [https://perma.cc/23TL-TUXP]; Cade Metz, *Artificial Intelligence Is Setting Up the Internet for a Huge Clash With Europe*, WIRED (July 11, 2016), https://www.wired.com/2016/07/artificial-intelligence-setting-internet-huge-clash-europe/ [https://perma.cc/GFY4-D4SR]; Ian Sample, *AI Watchdog Needed to Regulate Automated Decision-making, Say Experts*, GUARDIAN (Jan. 27, 2017), https://www.theguardian.com/technology/2017/jan/27/ai-artificial-intelligence-watchdog-needed-to-prevent-discriminatory-automated-decisions [https://perma.cc/J4KB-WVEL]; Matt Burgess, *Watching Them, Watching Us: Can We Trust Big Tech to Regulate Itself?*, CREATIVE REV., (Apr. 2017), https://www.creativereview.co.uk/watching-watching-us/ [https://perma.cc/85ZF-KWLA]; ACM U.S. Pub. Policy Council, Statement on Algorithmic Transparency and Accountability (May 25, 2017).

[93] *See* GDPR, *supra* note 19, at chs. 6, 8; Lomas, *supra* note 30 (noting that the "beefing up of enforcement that's baked into the new regime means there's a better opportunity for DPAs to start to bark and bite like proper watchdogs").

[94] *See* Lomas, *supra* note 30.

[95] In addition to the DPD, there are numerous other regulations that allude to rights involving automated decision-making explicability. "For example, the public sector is subject

rights that specifically address automated decision-making.[96] What these rules have lacked, however, is a meaningful threat of enforcement.[97]

Under the DPD, EU agencies tasked with carrying out its mandate were highly limited in their capacity to levy financial penalties against entities breaching the DPD.[98] Before the GDPR, the UK's Information Commissioner's Office (ICO), for example, was capped at a maximum fine of just £500,000 for violations.[99] Facebook's annual revenue for the 2017 fiscal year, by comparison, topped $40B.[100] Therefore, at most, the ICO could only hope to impose a fine representing a paltry percentage of the company's annual revenue.

Moreover, replacing the DPD with the GDPR represents an instance of an EU Regulation replacing a Directive. While directives "set out general rules to be transferred into national law by each country as they deem appropriate," regulations constitute a single, uniform law that is "directly applicable" to all

to the Public Administration Act that requires, *inter alia*, individual decisions to be substantiated. The person concerned has the right to be informed of the regulations and the actual circumstances underpinning a decision, as well as the main considerations that have been decisive." ARTIFICIAL INTELLIGENCE AND PRIVACY, *supra* note 4, at 22 (quoting Public Administration Act Sections 24 and 25). The EU also explicitly treats privacy protection as a fundamental right.

[96] *See* DPD, *supra* note 3; *see also, e.g.*, Mendoza & Bygrave, *supra* note 91; Lee A. Bygrave, *Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling*, 17 COMPUTER L. & SECURITY REV. 17 (2001); Alfred Kobsa, *Tailoring Privacy to Users' Needs*, *in* PROCEEDINGS OF THE 8TH INTERNATIONAL CONFERENCE ON USER MODELING 303 (M. Bauer et al. eds., 2001); Mireille Hildebrandt, *Profiling and the Rule of Law*, 1 IDENTITY IN INFO. SOC'Y 55, 55 (2008). Wachter, et al. actually discuss this phenomenon, noting: "Interestingly, despite years of negotiations, the final wording of the GDPR concerning protections against profiling and automated decision-making hardly changed from the relevant Articles and Recitals of the Data Protection Directive [of] 1995." Wachter et al., *supra* note 46, at 81. But their failure to address the enhanced enforcement powers introduced by the GDPR renders moot their underlying argument that the new provisions will do little to change the current regulatory landscape.

[97] *See* Mendoza & Bygrave, *supra* note 91, at 78 (describing art. 15 as "a second-class data protection right: it is rarely enforced, poorly understood and easily circumvented").

[98] *See id.*; DPD, *supra* note 3.

[99] *Facebook Faces £500,000 Fine from UK Data Watchdog*, BBC NEWS (July 11, 2018), https://www.bbc.com/news/technology-44785151 [https://perma.cc/P6E7-QNNB]. The GDPR specifies the monetary sanctions available to DPAs, unlike the DPD which left it to countries to set their own sanctions. *See* DPD, *supra* note 3, at art. 24 (leaving it to "Member States [to] adopt suitable measures to ensure the full implementation of the provisions of this Directive and shall in particular lay down the sanctions to be imposed in case of infringement of the provisions adopted pursuant to this Directive").

[100] *See* Press Release, Facebook Investor Relations, Facebook Reports Fourth Quarter and Full Year 2017 Results (Jan. 31, 2018) [hereinafter Facebook Press Release].

EU Member States.[101] The differences between these two paths to legislative implementation may seem trivial to outsiders looking in, but their practical effects are not. Unlike the GDPR, the DPD is subject to twenty-eight different interpretations and enforcement regimes—leading to differences that can foment confusion and inconsistency among industry leaders and data protection authorities alike. Coupled with the limited fines available under the DPD, these inconsistencies exacerbated enforcement problems for data protection authorities.

The combined effect of these DPD enforcement limitations produced a pack of EU data watchdogs tethered to a markedly short regulatory leash. For over two decades, the Directive set a high standard for data protection for companies handling the personal information of EU citizens. But those responsible of upholding these protections have long been perceived as lacking a genuine threat of enforcement.

Viewed through this lens, it is easy to understand why the debate surrounding the "right to explanation" has seen comparatively little attention paid to the authorities that will actually be tasked with enforcing it. For if the past were prologue, they could be expected to play a peripheral role in carrying out the right's protective mandate. However, with the passage of the GDPR, all of that is set to change. Chapters 6 and 8 of the Regulation grant data authorities vastly increased investigatory powers, an enhanced "enforcement tool kit," and the capacity to levy far greater financial penalties against entities in breach.[102]

EU data authorities will no longer be constrained by the limited range of enforcement options available under the DPD. Instead, these authorities will have far-reaching investigatory and corrective powers that allow them to issue sanctions against data protection violations that are "effective, proportionate," and, most importantly, "dissuasive."[103] Whereas data authorities under the DPD were limited to six-figure fines or sternly-worded letters, companies now will live under the threat of corrective measures that may be on orders of magnitude more potent.[104] Under this new reality, some commentators have

[101] KAREN DAVIES, UNDERSTANDING EUROPEAN UNION LAW (6th ed., 2016). Art. 288 of the Treaty on the Functioning of the European Union provides that: "A directive shall be binding, as to the result to be achieved, upon each Member State to which it is addressed, but shall leave to the national authorities the choice of form and methods." Consolidated Version of the Treaty on European Union art. 288, 2006 O.J. C 321 E/5. at 126. Article 288 states that a regulation, on the other hand, "shall be binding in its entirety and directly applicable in all Member States." *Id.* at 125.

[102] *See infra* Section III.A and accompanying notes.

[103] *See* GDPR, *supra* note 19, at art. 83.

[104] *See supra* note 99 and accompanying text.

asserted that the transition from the DPD to the GDPR should be understood as less about "individual EU Member States . . . getting stronger privacy laws" and more about EU data authorities finally starting "to bark and bite like proper watchdogs."[105]

The following subparts describe the specific enforcement powers that the GDPR provides European data authorities, as well as some of the practical implications of this power shift for downstream enterprises.

A. The Ascent of Enforcement

Chapter 6 of the GDPR provides for the appointment, by each Member State, of "one or more independent public authorities to be responsible for monitoring [its] application . . . ."[106] The legislation endows these agencies—which it terms "supervisory authorities" (SAs)—with broad "investigatory," "advisory," and "corrective" powers of far greater scope than those currently available under the DPD.[107] According to Chapter 6, these powers ensure the "consistent application" of the GDPR throughout the EU and include, among many other provisions, the ability: (1) "to obtain . . . access to all personal data [belonging to a company] and to all information necessary for the performance of [investigatory] tasks," (2) "to carry out investigations in the form of data protection audits,"[108] (3) "to issue warnings [or] reprimands to a [company]," (4) "to impose a temporary or definitive limitation [against companies] including a ban on processing," and (5) "to order the suspension of data flows to a recipient in a third country[109] or to an international organisation."[110]

Chapter 6's expansive set of investigatory and corrective powers are buttressed by an equally expansive set of remedial powers laid out in Chapter 8. These powers provide supervisory agencies with the authority to impose administrative fines that are "effective, proportionate, and dissuasive."[111] Under Chapter 8, SAs can fine companies that violate the GDPR's basic administrative or technical requirements up to €10 million or up to 2% of the companies' total annual revenue for the preceding financial year, "whichever

[105] Lomas, *supra* note 30.

[106] *See* GDPR, *supra* note 19, at art. 51.

[107] *See id.* at art. 58.

[108] Data protection audits are discussed in greater detail in Section IV.C *infra*.

[109] This term is discussed in detail in *infra* Part V.

[110] GDPR, *supra* note 19, at art. 58.

[111] *See* GDPR, *supra* note 19, at art. 83. The DPD, by contrast, places authority for adopting "suitable measures to ensure the full implementation of the provisions" with individual Member States. This has led to highly limited enforcement capabilities. DPD, *supra* note 3, at art. 24; *see supra* notes 99–100 and accompanying text.

is higher."[112] For violations of provisions more fundamental to the GDPR's data protection mandate[113]—including Articles 13, 14, 15, and 22—the maximum allowable fine increases precipitously. SAs can punish infringers of these provisions with fines of up to €20 million, or up to 4% of the companies' total annual revenue for the preceding financial year—again, "which[ever] is higher."[114]

The operative adjective, in both such instances, is the word "higher." To return to the example of the tech giant Facebook, whose annual revenues approximate €40 billion, a fine of 4% of annual turnover could total €1.6 billion, more than 3,200 times larger than the maximum fine available in the UK under the DPD.[115] This switch from proportional, as opposed to fixed, financial penalties ensures that even the titans of industry will not be immune from enforcement.

But for any in-house practitioners whose pulse doubled at the sight of such a multiple, the Regulation also provides cause for relief. First, the GDPR makes clear that punishment for breaches should be individualized and proportionate. The GDPR does not mandate the use of fines for all enforcement actions.[116] Article 83 outlines an extensive list of considerations for SAs seeking to ensure that their punishments are commensurate with the alleged violation.[117] These factors shift the administrative focus to the actual impacts of the violation, including the number of individuals affected, the actual damages suffered, and the sensitivity of the personal data at root.[118] Also, the GDPR stipulates that good faith efforts to proactively implement protective policies, ensure transparency, notify enforcement agencies, and cooperate with SA oversight will further reduce the likelihood of companies facing serious sanctions.[119]

[112] *See* GDPR, *supra* note 19, at art. 83.

[113] "Examples that fall under this category are non-adherence to the core principles of processing personal data, infringement of the rights of data subjects and the transfer of personal data to third countries or international organizations that do not ensure an adequate level of data protection." *GDPR: Guidelines and Consequences for Non-Compliance*, GDPR:REPORT (June 16, 2017), https://gdpr.report/news/2017/06/16/gdpr-guidelines-consequences-non-compliance/ [https://perma.cc/J756-M5XD]. *See* GDPR, *supra* note 19, at art. 84.

[114] *See* GDPR, *supra* note 19, at art. 83.

[115] *See supra* notes 99–100 and accompanying text.

[116] *See* GDPR, *supra* note 19, at art. 83.

[117] *See id.*

[118] *See id.*

[119] *See id.*

B. THE IMPORTANCE OF UNDERSTANDING WHEN THE WATCHDOGS MIGHT BITE

With great power, of course, comes great interpretive responsibility. After all, what better source of guidance could there be for companies seeking to ensure compliance with the GDPR's "right to explanation" than the data authorities likeliest to bring enforcement action against them? Any agency action will, of course, be subject to the slower-burning process of judicial clarification through national and international litigation. But while any such activity percolates through the EU's multi-layered legal system, the de facto interpretive authorities of the "right to explanation" will be those whose primary responsibility it is to investigate and punish companies that breach the GDPR.

Data protection authorities have already begun to signal their anticipated ascendance by flexing additional regulatory muscle in the lead up to the GDPR's effectuation.[120] According to a recent report, the total monetary value of fines the UK's ICO levied doubled in 2016—coinciding with a steep uptick in the number of enforcement notices issued by the agency and a nearly 100% increase in the size of its fines.[121] This increased enforcement activity also came amid calls by the agency to increase its staff size in advance of the GDPR's May 2018 effectuation.[122]

[120] *See* Max Metzger, *Sharp Rise in ICO Fines and Enforcement Notices as GDPR Races Closer*, SC MEDIA (June 1, 2017), https://www.scmagazineuk.com/sharp-rise-in-ico-fines-and-enforcement-notices-as-gdpr-races-closer/article/665466/ [https://perma.cc/PT6D-EXMU]; Elizabeth Denham, the residing commissioner, remarked:

> In this world of big data, AI and machine learning, my office is more relevant than ever. I oversee legislation that demands fair, accurate and non-discriminatory use of personal data; legislation that also gives me the power to conduct audits, order corrective action and issue monetary penalties. Furthermore, under the GDPR my office will be working hard to improve standards in the use of personal data through the implementation of privacy seals and certification schemes. We're uniquely placed to provide the right framework for the regulation of big data, AI and machine learning, and I strongly believe that our efficient, joined-up and co-regulatory approach is exactly what is needed to pull back the curtain in this space.

Elizabeth Denham, *Information Commissioner's Foreword*, *in* BIG DATA, ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND DATA PROTECTION 3 (2017); *see also* Jamie Doward et al., *Watchdog to Launch Inquiry into Misuse of Data in Politics*, GUARDIAN (Mar. 4, 2017), https://www.theguardian.com/technology/2017/mar/04/cambridge-analytics-data-brexit-trump [https://perma.cc/B3ST-5H5H].

[121] *See* Metzger, *supra* note 120.

[122] *See id.*

IV. THE NEXT CHAPTER IN THE DEBATE: SA ENFORCEMENT AND THE RISE OF DATA AUDITS

Viewed against the backdrop of Chapter 6's and 8's vastly enhanced enforcement powers, it becomes immediately apparent that the public debate over the "right to explanation" can no longer be confined exclusively to the text of the GDPR. Instead, the right articulated by the Regulation must be understood holistically with a newfound deference owed to the downstream interpretations by the EU data watchdogs whose regulatory bark and bite will soon become far costlier for companies to ignore. Fortunately, a recent burst of activity by these very data authorities has provided extensive guidance for enterprises seeking to better understand what meaningful compliance with the GDPR's controversial "right to explanation" entails in practice.

The following subparts detail these new activities, relying on the words of the data authorities themselves whenever possible in order to minimize the likelihood of editorializing. Subpart A details the recent activity by the Article 29 Data Protection Working Party, a European body charged with a senior advisory role in the GDPR's implementation. Subpart B then takes the interpretation of a single data protection authority, the UK's Information Commissioner's Office (ICO), as a case study for understanding the scope of the "right to explanation" in practice.

A. THE INTERPRETATION OF THE ARTICLE 29 DATA PROTECTION WORKING PARTY

In October 2017, the Article 29 Data Protection Working Party (A29WP) published its official "Guidelines on Automated Individual Decision-Making and Profiling" for the GDPR.[123] The A29WP "is the European Commission's most senior advisory body on data protection and information security matters" and serves as a central authority for all EU data protection agencies.[124] Although its guidelines are nonbinding, they constitute a vital reference point for the individual SAs appointed by EU Member States and are, therefore,

[123] *See* A29WP Automated Decision-Making Guidelines, *supra* note 46.

[124] ARTIFICIAL INTELLIGENCE AND PRIVACY, *supra* note 4, at 4. The A29WP, which launched in 1996, derives its name from Article 29 of the DPD setting out its composition and purpose. *See Glossary A*, EUROPEAN DATA PROTECTION SUPERVISOR, https://edps.europa.eu/data-protection/data-protection/glossary/a_en [https://perma.cc/3CTD-8T8E] (noting the " 'Article 29 Working Party' is the short name of the Data Protection Working Party established by Article 29 of Directive 95/46/EC"). It is a representative body composed of data protection authorities from each EU Member State, and it also includes the European Data Protection Supervisor and the European Commission. Since the GDPR took effect, it has been replaced by the "European Data Protection Board." *See* GDPR *supra* note 19, at art. 68.

critical to understanding how those authorities should interpret the GDPR.

The A29WP's guidance on automated decision-making included numerous provisions intended to clarify the "right to explanation"—stemming from a collection of rights that the A29WP referred to as the rights "to be informed," "to obtain human intervention," and "to challenge [a] decision" made by certain automated systems.[125] According to the A29WP, the "complexity of machine-learning" algorithms used in such systems "can make it challenging to understand how an automated decision-making process or profiling works."[126] But such complexity, it insisted, "is no excuse for failing to provide information" to data subjects.[127] The A29WP instructed that companies making automated decisions that fall under Article 22(1) "should find simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision"—albeit "without necessarily always attempting a complex explanation of the algorithms used or [a] disclosure of the full algorithm."[128] In doing so, the A29WP stipulated that companies must:

- "[T]ell the data subject that they are engaging in this type of activity;
- [P]rovide meaningful information about the logic involved; and
- [E]xplain the significance and envisaged consequences of the processing."[129]

The A29WP further clarified that the phrase "[m]eaningful information about the logic involved will in most cases require controllers to provide details such as":

- "[T]he information used in the automated decision-making process, including the categories of data used in a profile;
- [T]he source of that information;
- [H]ow any profile used in the automated decision-making process is built, including any statistics used in the analysis;
- [W]hy this profile is relevant to the automated decision-making process; and

[125] *See* A29WP Automated Decision-Making Guidelines, *supra* note 46, at 9.

[126] *Id.* at 14.

[127] *Id.* at 14 n.12.

[128] *Id.* at 14.

[129] *Id.* at 13–14.

- [H]ow it is used for a decision concerning the data subject."[130]

The A29WP added that it was "good practice [for companies] to provide the above information *whether or not* the processing falls within the narrow Article 22(1) definition."[131] The agency also insisted that companies could not avoid Article 22 by simply "fabricating" *de minimus* human involvement in decision-making.[132] According to the A29WP, companies must ensure that any human "oversight of [a] decision is meaningful, rather than just a token gesture" if they intend for their systems to fall outside the scope of Article 22's provisions pertaining to decisions "based *solely* on automated processing."[133]

In addition to the specific explanatory measures outlined above, the A29WP also recommended that companies introduce more general "procedures and measures to prevent errors, inaccuracies or discrimination" in data processing.[134] The guidelines suggested that companies "carry out frequent assessments on the data sets they process to check for any bias, and develop ways to address any prejudicial elements, including any over-reliance on correlations."[135] According to the A29WP, these assessments should be conducted "on a cyclical basis; not only at the design stage, but also continuously, as the profiling is applied to individuals," so that the "outcome of such testing [can] feed back into the system design."[136]

One such safeguard the A29WP repeatedly invoked involves the use of the "Data Protection Impact Assessment" (DPIA), originating under Article 35 of

[130] *Id.* at 28.

[131] *Id.* at 13 (emphasis added). This justification stemmed, in part, from GDPR Recital 60 stating:

> The controller should provide the data subject with any further information necessary to ensure fair and transparent processing taking into account the specific circumstances and context in which the personal data are processed. Furthermore, the data subject should be informed of the existence of profiling and the consequences of such profiling.

GDPR *supra* note 19, at Recital 60.

[132] A29WP Automated Decision-Making Guidelines, *supra* note 46, at 10.

[133] *Id.* at 13 (emphasis added). This question, too, has been the subject of heated debate due to Article 22's use of the phrase "solely" in its provisions related to automated decision-making. *See, e.g.*, Wachter et al., *supra* note 46, at 88; Selbst & Powles, *supra* note 46, at 5–6. The A29WP further clarified that: "[i]t should be carried out by someone who has the authority and competence to change the decision. As part of the analysis, they should consider all the available input and output data." A29WP Automated Decision-Making Guidelines, *supra* note 46, at 10.

[134] A29WP Automated Decision-Making Guidelines, *supra* note 46, at 17.

[135] *Id.*

[136] *Id.*

the GDPR.[137] Although the GDPR does not formally define the concept of the DPIA, the A29WP described it as "a process for building and demonstrating" compliance by systematically examining automated processing techniques to determine the measures necessary to "manage the risks to the rights and freedoms of natural persons resulting from the processing of personal data."[138]

While noting that the GDPR provides companies with considerable "flexibility to determine the precise structure and form of the DPIA," the A29WP stipulated that the DPIA represented a fundamentally "iterative process" with "common criteria" for carrying it out.[139] According to the A29WP, these criteria were best understood as falling within the GDPR's broader "data protection by design" principles, which apply at all stages of a system's life cycle.[140]

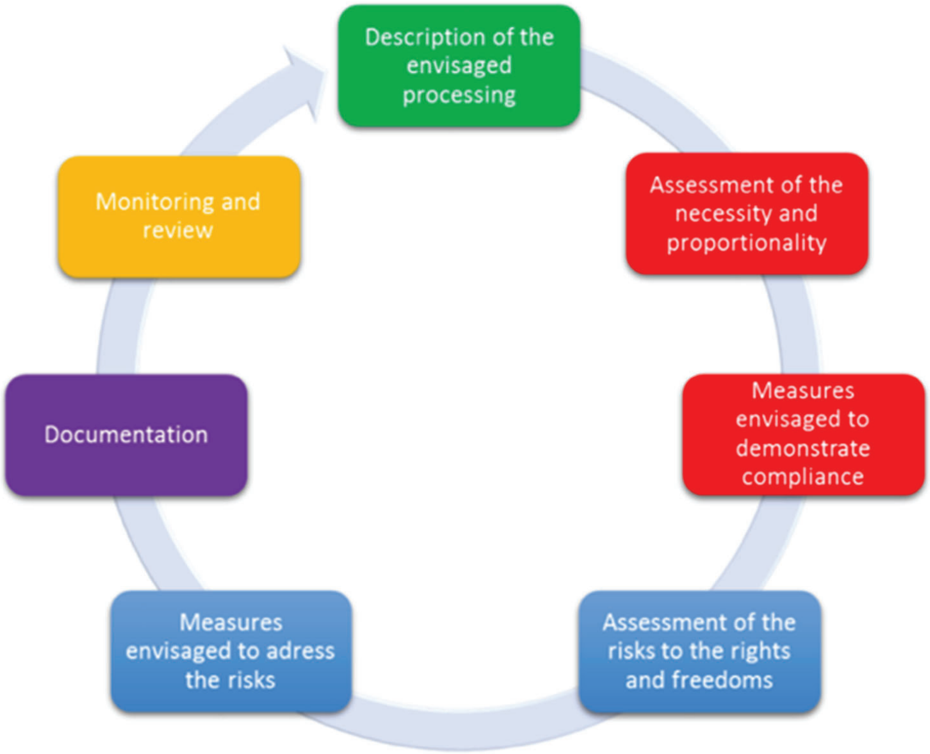
[137] *Id.* at 27.

[138] *See* ARTICLE 29 WORKING PARTY, GUIDELINES ON DATA PROTECTION IMPACT ASSESSMENT (DPIA) AND DETERMINING WHETHER PROCESSING IS "LIKELY TO RESULT IN A HIGH RISK" FOR THE PURPOSES OF REGULATION 2016/679 4 (2017) [hereinafter A29WP DPIA Guidelines].

[139] The A29WP DPIA Guidelines Annexes 1 and 2 provide additional details regarding these requirements. *See id.* at 21–22.

[140] *See id.* at 14.

Figure I: The Iterative DPIA Process[141]


Description of the envisaged processing
Assessment of the necessity and proportionality
Measures envisaged to demonstrate compliance
Assessment of the risks to the rights and freedoms
Measures envisaged to adress the risks
Documentation
Monitoring and review

Under the GDPR's "data protection by design" mandate, companies must "[t]ak[e] into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by [] processing."[142] The GDPR recommends DPIAs as a means of proactively identifying and addressing these considerations so that companies can effectively "implement appropriate technical and organisational . . . safeguards into the[ir] processing [operations]."[143]

1. *When Are DPIAs More Than Mere Recommendations?*

The A29WP's guidance stresses that, in many circumstances, DPIAs are not merely recommended as a matter of best practices but are compulsory. In

[141] *Id.* at 16.

[142] GDPR, *supra* note 19, at art. 25.

[143] *Id.* The GDPR explicitly recommends "measures, such as pseudonymisation, which are designed to implement data-protection principles, [and] data minimisation." *Id.*

determining whether a DPIA is or is not compulsory, Article 35(1) of the GDPR relies, primarily, on the heuristic of so-called "high risk" data processing operations.[144] According to the Regulation, DPIAs are mandatory "[w]here a type of processing . . . taking into account the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons . . . ."[145] Article 35 establishes a non-exhaustive list of scenarios likely to be deemed high risk, including when operations involve:

> a) [A] systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person;
>
> b) [P]rocessing on a large scale of special categories of data referred to in Article 9(1),[146] or of personal data relating to criminal convictions and offences referred to in Article 10;[147] or
>
> c) [A] systematic monitoring of a publicly accessible area on a large scale.[148]

The A29WP's guidance elaborates on this list by enumerating ten specific scenarios that "provide a more concrete" set of criteria for determining

[144] *See id.* at art. 35.

[145] *Id.*

[146] *See id.* Article 9(1) states:

> Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.

Id. at art. 9.

[147] *See id.* at art. 35. Article 10 states:

> Processing of personal data relating to criminal convictions and offences or related security measures based on Article 6(1) shall be carried out only under the control of official authority or when the processing is authorised by Union or Member State law providing for appropriate safeguards for the rights and freedoms of data subjects. Any comprehensive register of criminal convictions shall be kept only under the control of official authority.

Id. at art. 10. Article 6(1) includes a list of criteria for establishing the lawfulness of processing. *See id.* at art. 6(1).

[148] *Id.* at art. 35. The GDPR notes that the use of "new technologies" is "particularly" likely to produce high risks. *See id.*

whether operations are "high risk." These include instances where processing involves: (1) evaluating or scoring, (2) automated decision-making with legal or similarly significant effects, (3) systematic monitoring, (4) sensitive data, (5) data processed on a large scale, (6) datasets that have been matched or combined, (7) data concerning vulnerable data subjects, (8) innovative use or applying technological or organizational solutions, (9) data transfer across borders outside the European Union, and (10) processing that inherently "prevents data subjects from exercising a right or using a service or a contract."[149]

Although the A29WP emphasized that DPIAs are not obligatory "for every processing operation which may result in risks," the GDPR's requirement that an *ex ante* assessment be conducted for all processing operations produces a distinctly circular effect.[150] In cases where it is unclear whether a given operation requires a DPIA, carrying out a preliminary DPIA to assess the risks may be the best means of ensuring compliance. In other words, demonstrating that a DPIA is not necessary will, in many instances, itself require a DPIA.[151] This somewhat circular effect will likely incentivize companies to err on the side of caution with DPIAs. Companies may implement them even if the intent in doing so is to simply document or investigate whether more robust explanatory measures are required.

Crucially, these *ex ante* assessments are required even when the GDPR's provisions pertaining to decision-making "based *solely* on automated processing" are not directly implicated.[152] The A29WP repeatedly highlighted that Article 35(3)(a)'s deliberate exclusion of the word "solely" meant that the Article "appl[ied] in the case of decision-making including profiling with legal or similarly significant effects that is *not wholly automated*, as well as solely automated decision-making defined in Article 22(1)."[153]

[149] *See* A29WP DPIA Guidelines, *supra* note 138, at 9–11.

[150] *See id.* at 8.

[151] The A29WP DPIA Guidelines stressed that:

> In order to enhance compliance with this Regulation where processing operations are likely to result in a high risk to the rights and freedoms of natural persons, the controller should be responsible for the carrying-out of a data protection impact assessment to evaluate, in particular, the origin, nature, particularity and severity of that risk.

Id. at 4.

[152] *See* A29WP Automated Decision-Making Guidelines, *supra* note 46, at 10 (emphasis added).

[153] *See id.* at 29 (emphasis added).

2. *What Kinds of Documented Explanations Do DPIAs Require?*

As a means of promoting additional transparency through DPIAs, the A29WP instructed that when data "processing is wholly or partly performed by a [company]," the company should assist SAs "in carrying out [a] DPIA and provide any necessary information" to them.[154] Moreover, the A29WP emphasized that, under Article 35(9), companies are required, "where appropriate," to actively "seek the views of data subjects or their representatives" during the DPIA process.[155] In fulfilling this obligation, the A29WP stated that the views of data subjects could be solicited by a variety of means "depending on the context," including "an internal or external study related to the purpose and means of the processing operation," "a formal question" directed to the relevant stakeholders, or "a survey sent to the data controller's future customers."[156] The A29WP also noted that when a company's "final decision" to proceed with a particular process operation "differ[ed] from the views of the data subjects, its reasons for going ahead or not should be [also] documented."[157] Even in instances where a company has decided that soliciting the views of data subjects is not appropriate, the A29WP insisted that the company should nonetheless document "its justification for not seeking the views of data subjects."[158]

Article 35(7) of the GDPR specifically enumerates four basic features that all DPIAs must, at a minimum, contain:

> 1. [A] systematic description of the envisaged processing operations and the purposes of the processing, including, where applicable, the legitimate interest pursued by the controller;
>
> 2. [A]n assessment of the necessity and proportionality of the processing operations in relation to the purposes;
>
> 3. [A]n assessment of the risks to the rights and freedoms of data subjects[; and]
>
> 4. [T]he measures envisaged to address the risks, including safeguards, security measures and mechanisms to ensure the

[154] A29WP DPIA Guidelines, *supra* note 138, at 15.

[155] *Id.*

[156] Damiana Lesce, Paola Lonigro & Valeria de Lucia, *Privacy. Data Protection Impact Assessment (DPIA). The Art. 29 Data Protection Working Party Guidelines*, LEXOLOGY, https://www.lexology.com/library/detail.aspx?g=b7e8d97f-dd45-48de-8796-c68c2e5bf0a9 [https://perma.cc/WC7L-HWFB].

[157] A29WP DPIA Guidelines, *supra* note 138, at 15.

[158] *Id.*

> protection of personal data and to demonstrate compliance with this Regulation taking into account the rights and legitimate interests of data subjects and other persons concerned.[159]

Finally, the A29WP added that while publicly releasing "a DPIA is not a legal requirement of the GDPR," companies "should consider publishing . . . their DPIA[s]" either in full or in part.[160] The A29WP stated that the "purpose of such a process would be to help foster trust in the controller's processing operations, and demonstrate accountability and transparency"—particularly "where members of the public are affected by the processing operation."[161] According to the institution, the "published DPIA does not need to contain the whole assessment, especially when the DPIA could present specific information concerning security risks for the data controller or give away trade secrets or commercially sensitive information" and "could even consist of just a summary of the DPIA's main findings."[162]

B. From the A29WP to Supervisory Authorities

From the central guidance provided by the A29WP come the specific downstream interpretations of EU data authorities. Although the individual interpretations of these SAs are, by design, the furthest from the textual wellspring of the GDPR, they are by far the most relevant for companies seeking to promote compliance. As the agencies on the front lines of overseeing investigations and issuing sanctions, the interpretations they provide will constitute the clearest signals for companies attempting to understand the substantive protections afforded by the GDPR's "right to explanation."

1. *Why the ICO?*

The analysis that follows focuses on one such authority—the UK's Information Commissioner's Office (ICO). The reasons for this focus on the ICO are twofold. First, surveying all twenty-eight agencies would be needlessly exhaustive, as each agency's interpretation draws directly from the GDPR as opposed to drawing indirectly from twenty-eight individual legislative enactments, as was the case under the DPD. Second, and most importantly, the UK's imminent exit from the EU makes the ICO a particularly informative example. Despite the imminent separation from the European bloc, the country seeks to continue the free flow of data with Continental Europe by

[159] *Id* at 4.

[160] *Id* at 18.

[161] *Id.*

[162] *Id.*

promoting domestic compliance with the GDPR. Thus, the fact that the ICO is, in one sense, a bad example makes it an especially good one. The agency, after all, will be particularly attuned to ensuring its framework is coextensive with the rest of the EU's.

2. *The ICO's Guidance*

Since the A29WP's release of its GDPR guidance in October 2017, the ICO, along with every other EU data authority, published extensive guidelines for organizations seeking to comply with the GDPR's requirements.[163] The agency describes these guidelines as a "living document" subject to elaboration or alteration on an ongoing basis.[164] Among the ICO's many provisions interpreting the GDPR are those pertaining to the data subjects' "rights related to automated decision making including profiling."[165] According to the ICO, companies processing data "must identify whether any of [their] processing falls under Article 22 and, if so, make sure that" they:

- "[G]ive individuals information about the processing;
- [I]ntroduce simple ways for them to request human intervention or challenge a decision;
- [C]arry out regular checks to make sure that your systems are working as intended."[166]

When processing operations fall under Article 22's specific purview,[167] the ICO also requires that companies carry out a DPIA "to identify the risks to individuals," to "show how [they] are going to deal with them," and to

[163] *See generally Guide to the General Data Protection Regulation (GDPR)*, INFO. COMMISSIONER'S OFF., https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/ [https://perma.cc/2GC6-4GEC] (last visited Apr. 2, 2019). The UK Government has also issued new data protection legislation that will implement the standards set forth by the GDPR. *See GDPR Fact Sheet*, BENEFACTO, https://benefacto.org/gdpr-fact-sheet/ [https://perma.cc/5KJD-T8C3] (last visited Apr. 2, 2019). These laws include a number of additional protections going above and beyond the baseline set by the GDPR which extend to "journalists, scientific and historical researchers, and anti-doping agencies who handle people's personal information." *Id.*

[164] *See* ICO'S OVERVIEW OF GDPR, *supra* note 14, at 3.

[165] *See Rights Related to Automated Decision Making Including Profiling*, INFO. COMMISSIONER'S OFF., https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/rights-related-to-automated-decision-making-including-profiling/ [https://perma.cc/6SEH-DNKD] [hereinafter ICO Automated Decision Making Guidelines].

[166] *Id.* Notably, this mandate is coextensive with the A29WP's own non-binding recommendation, which the ICO appears to be diligently replicating.

[167] *See id.* Some instances do not apply. *See supra* Part II.A.

demonstrate the "measures [they] have in place to meet GDPR requirements."[168]

Even when processing operations fall outside of Article 22, the ICO's guidelines explicitly endorse the use of a DPIA as part of a broader compliance tool kit based on the same principles of "data protection by design" (DPbD) identified by the A29WP.[169] In addition to the comprehensive set of recommendations involving DPbD detailed in its public discussion paper,[170] the ICO states that companies "have a general obligation to implement technical and organisational measures to show that [they] have considered and integrated data protection into [their] processing activities."[171]

C. The Rise of the DPIA and Data Protection by Design

From the guidance set forth by the A29WP and the ICO, one fact is overwhelmingly clear: the GDPR's "right to explanation" is no mere remedial mechanism to be invoked by data subjects on an individual basis, but it implies a more general form of oversight with broad implications for the design, prototyping, field testing, and deployment of data processing systems. The "right to explanation" may not require that companies pry open their "black boxes" per se, but it does require that they evaluate the interests of relevant stakeholders, understand how their systems process data, and establish policies for documenting and justifying key design features throughout a system's life cycle. Not only must companies convey many of these details directly to downstream data subjects,[172] but they must also document and explain the safeguards in place for managing data processing risks either through a DPIA as described in Article 35 or through a substantively similar mechanism. Indeed, it is perhaps no coincidence that the formulation of Article 35(1) bears such a striking similarity to that of Article 22(1). Taken together, these two mandates produce a powerful synergistic effect that promotes the kinds of prophylactic DPbD principles prevalent throughout the GDPR.[173] As a

[168] *Id.* Even in instances where Article 22's requirements do not apply, the ICO recommends that companies nonetheless "carry out a DPIA to consider and address the risks before [they] start any new automated decision-making or profiling" and "tell [] customers about the profiling and automated decision-making [they] carry out, what information [they] use to create the profiles and where [they] get this information from." *Id.*

[169] *See Data Protection by Design and Default*, INFO. COMMISSIONER'S OFF., https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-by-design-and-default/ [https://perma.cc/E9FS-J4NM] (last visited Apr. 2, 2019).

[170] *See* ICO'S OVERVIEW OF GDPR, *supra* note 14, at 32–37.

[171] *Id.* at 32.

[172] *See supra* Section IV.A.

[173] *See* GDPR, *supra* note 19, at art. 25, Recital 78.

consequence, it now appears that *ex ante* DPIAs—as opposed to *ex post* invocations of an individual "right to explanation"—are destined to "become the required norm for algorithmic systems, especially where sensitive personal data, such as race or political opinion, is processed on a large scale."[174]

The advantages of shifting the dialogue surrounding the GDPR's "right to explanation" from one involving individual remedies to one involving more general DPbD principles are manifold. First, mere algorithmic explicability is not the panacea it is often presumed to be.[175] As numerous experts of diverse backgrounds have noted, the reliance on transparency as an individualized mechanism often places excessive burdens on resource-constrained users to "seek out information about a system, interpret it, and determine its significance, only then to find out they have little power to change things anyway, being disconnected from power."[176] Though transparency may often feel like a robust solution intuitively, explainable artificial intelligence—or

[174] *See* Edwards & Veale, *supra* note 73, at 78 (quoting GDPR, art. 35(3)(b)) (internal quotations omitted) (arguing that DPIAs will soon become mainstream in enterprise); *see also, e.g.*, A29WP DPIA Guidelines, *supra* note 138. This prediction involving the rise of data auditing methodologies is also supported by additional legal mechanisms within the GDPR that, for purposes of concision, are not addressed by this Article. *See, e.g.*, GDPR, *supra* note 19, at art. 42 (requiring "the establishment of data protection certification mechanisms and of data protection seals and marks . . . available via a process that is transparent" and subject to regular review); *id.* at art. 40 (recommending that companies "prepare codes of conduct . . . such as with regard to . . . fair and transparent processing" and "to carry out the mandatory monitoring of compliance").

[175] *But see* Kasper Lippert-Rasmussen, *"We Are All Different": Statistical Discrimination and the Right to Be Treated as an Individual*, 15 J. ETHICS 47, 54 (2011)

> [O]btaining information is costly, so it is morally justified, all things considered, to treat people on the basis of statistical generalizations even though one knows that, in effect, this will mean that one will treat some people in ways, for better or worse, that they do not deserve to be treated.

See, e.g., Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018) (describing increasingly vocal pushes for transparency due to the intuitive, but not always correct notion, that explanations will resolve unfairness within algorithms).

[176] *See* Edwards & Veale, *supra* note 73, at 67 (quoting Mike Annany & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, NEW MEDIA & SOC'Y 1, 5 (2018)) (internal quotations omitted); *see also, e.g.*, FRANK PASQUALE, *supra* note 68 (arguing that transparency in and of itself does not translate to accountability in many contexts); Joshua Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 638 (2017) (rejecting transparency as a true remedy for promoting accountability); Brendan Van Alsenoy et al., *Privacy Notices Versus Informational Self-Determination: Minding The Gap*, 28 INT'L REV. L., COMPUTERS & TECH. 185, 185 (2014) (arguing that privacy notices don't necessarily achieve the accountability goals that many expect they will).

"XAI"[177] as it is increasingly called—is especially unlikely to provide significant remedial utility to individuals in instances where the discrimination involved is only observable at the statistical scale. Moreover, some commentators have convincingly argued that too great a focus on individualized explanations—as opposed to broader, multi-methodological design practices for mitigating unfairness—could "nurture a new kind of transparency fallacy . . . ."[178] Indeed, providing a basic explanation to individual users could provide false cover for companies whose processing operations may be biased for other reasons.

Second, providing enterprises a broader range of compliance options could allow them greater flexibility when deploying machine learning systems that may make more conventional forms of explicability impractical or impossible.[179] Under the current state of the art, many of the highest performing machine learning algorithms pose significant "tradeoff[s] between the representational capacity of a model and its interpretability."[180] Techniques capable of achieving the richest predictive results tend to do so through the use of aggregation, averaging, or multilayered techniques which, in turn, make it difficult to determine the exact features that play the largest predictive role.[181] Depending on the circumstances, performance losses associated with adopting a more explicable approach could prove far costlier than the social utility of providing individualized explanations.[182] Particularly in instances where the leading techniques far outpace the remedial options available to data subjects, a one-size-fits-all approach to oversight could lead to unnecessary bureaucratic

[177] *See* Tim Miller, Explanation in Artificial Intelligence: Insights from the Social Sciences (June 22, 2017) (unpublished manuscript).

[178] *See* Edwards & Veale, *supra* note 73, at 81 (internal quotations omitted); *see also, e.g.*, Toon Calders & Indrė Žliobaitė, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 43, 46 (2013) ("[T]he selection of attributes by which people are described in [a] database may be incomplete.").

[179] *See supra* notes 68–71 and accompanying text.

[180] *See* Goodman & Flaxman, *supra* note 46, at 6. "Representational capacity" here refers, roughly, to the ability of an algorithm to make predictions that account for complex patterns, phenomenon, or inputs. Machine learning systems, especially those using deep neural networks, can give rise to models so complex that humans are unable to understanding precisely how the system arrives at a given decision or prediction.

[181] *See* Wojciech Samek et al., *Evaluating the Visualization of What a Deep Neural Network Has Learned*, 28 IEEE TRANSACTIONS ON NEURAL NETWORKS & LEARNING SYS. 2660, 2666–67 (2017); Marco Tulio Ribeiro et al., *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1135 (2016); Jon Kleinberg et al., *Human Decisions and Machine Predictions* (Nat'l Bureau of Econ. Research, Working Paper No. 23180, 2017).

[182] This, however, may eventually prove to be a moving target.

roadblocks for technologies with massively beneficial social impacts.[183]

Finally, and perhaps most importantly, system-wide audits of the type envisioned by DPIAs already have a well-documented track record of detecting and combating algorithmic discrimination in otherwise opaque systems. As Sandvig et al. note, audit studies are "the most prevalent social scientific methods for the detection of discrimination" in complex computational systems.[184] In recent years, these auditing techniques have been used by researchers and journalists to successfully detect and document algorithmic bias across diverse industry sectors and social domains.[185] Further, this approach includes the added benefit of allowing outside entities that may have more resources than individuals to scrutinize the integrity of complex computational systems. Regulators, NGOs, media outlets, and public interest organizations that specialize in this area will be able to invest in the expertise

[183] *See, e.g.*, Toon Calders & Sicco Verwer, *Three Naive Bayes Approaches for Discrimination-Free Classification*, 21 DATA MINING & KNOWLEDGE DISCOVERY 277 (2010) (describing trade-off between discrimination removal and classifier performance); Faisal Kamiran & Toon Calders, *Data Preprocessing Techniques for Classification Without Discrimination*, 33 KNOWLEDGE & INFO. SYS. 1 (2012) (describing trade-off between discrimination removal and classifier performance); Jagriti Singh & S. S. Sane, *Preprocessing Technique for Discrimination Prevention in Data Mining*, 4 INT'L J. ENGINEERING RES. & APPLICATIONS 54 (2014) (noting inherent trade-offs in the current state-of-the-art); Sam Corbett-Davies et al., Algorithmic Decision Making and the Cost of Fairness (June 2017) (unpublished manuscript). These tradeoffs will likely be a moving target. Indeed, Edwards & Veale note that the inevitability of these tradeoffs may only be "an interim conclusion" and are "convinced that recent research in ML explanations shows promise" for reducing or eliminating some of these tradeoffs. *See* Edwards & Veale, *supra* note 73, at 81.

[184] Christian Sandvig et al., Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms 5, 16 (May 22, 2014) (unpublished manuscript) (noting that the "audit study" is "the most prevalent social scientific method for the detection of discrimination" and that it is "considered to be the most rigorous way to test for discrimination in housing and employment"); Andrea Romei & Salvatore Ruggieri, *Discrimination Data Analysis: A Multi-Disciplinary Bibliography*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 109, 120 (2013); Faisal Kamiran, Toon Calders & Mykola Pechenizkiy, *Techniques for Discrimination Free Predictive Models*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 223, 223–24 (2013).

[185] *See generally, e.g.*, James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164, 173 (2017); FRANK PASQUALE, *supra* note 68; Mireille Hildebrandt, *The New Imbroglio - Living with Machine Algorithms*, *in* THE ART OF ETHICS IN THE INFORMATION SOCIETY 55 (Liisa Janssens ed., 2016); Kiel Brennan-Marquez, *"Plausible Cause": Explanatory Standards in the Age of Powerful Machines*, 70 VAND. L. REV. 1249, 1287 (2017); Andrew D. Selbst, *A Mild Defense of Our New Machine Overlords*, 70 VAND. L. REV. EN BANC 87 (2017); Reuben Binns, *Algorithmic Accountability and Public Reason*, 31 PHIL. & TECH. 543 (2018); Katherine Strandburg, N.Y. Univ. School of Law, Presentation at The Human Use of Machine Learning: An Interdisciplinary Workshop, Venice: Decision-Making, Machine Learning and the Value of Explanation (Dec. 16, 2016).

necessary not only to provide data subjects with the right answers but also to ensure that the right questions are asked.

Although data audit and DPbD methodologies come with their own unique set of challenges,[186] the multifaceted advantages[187] offered by these approaches present exciting new possibilities for fostering genuine algorithmic accountability in enterprises without stifling technological and business advances.[188] In contrast to a remedial "right to explanation" invoked on an individual basis by downstream data subjects, properly implemented auditing and DPbD can provide the evidence necessary to inform and vet the design and deployment of more fair, accountable, and transparent algorithmic systems.[189]

V. EXPORTING THE "RIGHT TO EXPLANATION": THE BRUSSELS EFFECT AND THE GDPR'S LONG TENTACLES

Although the EU is sometimes maligned as a declining force on the world stage, numerous recent studies have demonstrated that it actually exercises "unprecedented global power . . . through its legal institutions and standards that it successfully exports to the rest of the world . . . ."[190] This "export" effect

[186] *See* Bryce Goodman, A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection 7 (2017) (unpublished manuscript)

> [A] process that passes a safety audit may fail for other reasons (e.g., inefficiency). Passing a safety audit does not mean that all risk is eliminated but, rather, that risk is reduced to an acceptable level. Choosing an acceptable level of risk depends in turn on the process evaluated and, in particular, both the likelihood and severity of a failure.

See also Lior Jacob Strahilevitz, *Privacy Versus Antidiscrimination*, 75 U. CHI. L. REV. 363, 364 (2008).

[187] The list enumerated above is, of necessity, far from exhaustive.

[188] *See* Goodman, *supra* note 186, at 7.

[189] *See id.*; *see also* Anupam Datta et al., *Algorithmic Transparency via Quantitative Input Influence*, *in* TRANSPARENT DATA MINING FOR BIG AND SMALL DATA 71, 87–89 (Tania Cerquitelli et al. eds., Springer 2017).

[190] Anu Bradford, *The Brussels Effect*, 107 NW. U. L. REV. 1, 64 (2012); *see* Case COMP/M.5984, Intel/McAfee, SG-Greffe (2011) D/1407, C(2011) 529, EUR-Lex 32011M5984 (Jan. 26, 2011); *see also, e.g.*, Christopher Kuner, *The Internet and the Global Reach of EU Law* (LSE Legal Studies, Working Papers No. 4/2017, 2017); David Scheer, *Europe's New High-Tech Role: Playing Privacy Cop to the World*, WALL ST. J. (Oct. 10, 2003), https://www.wsj.com/articles/SB106574949477122300 [https://perma.cc/9LZK-XCZB]; Brandon Mitchener, *Rules, Regulations of Global Economy Are Increasingly Being Set in Brussels*, WALL ST. J. (Apr. 23, 2002), https://www.wsj.com/articles/SB1019521240262845360 [https://perma.cc/J8MS-DREP]; *Regulatory Imperialism*, WALL ST. J. (Oct. 26, 2007),

occurs through the process of "unilateral regulatory globalization." This entails a process whereby "a single state is able to externalize its laws and regulations outside its borders through market mechanisms, resulting in the globalization of standards."[191] Particularly in the last decades, the EU has evinced "a strong and growing ability to promulgate regulations that become entrenched in the legal frameworks of developed and developing markets alike" without relying on international institutions or intergovernmental negotiations.[192] This phenomenon has since come to be described as the "Brussels Effect."[193]

The following subparts explore this effect on enterprises seeking to comply with the EU's data protection mandate. Section A describes the DPD's influence as a global "gold standard" since 1995 as well as the potential consequences of this phenomenon for the GDPR's own global legacy. Section B then details the implications of the GDPR's "Brussels Effect" for individual enterprises and concludes by documenting some of the real-world impacts technology companies have already experienced.

A. DATA PROTECTION AND THE "BRUSSELS EFFECT"

There is, perhaps, no better exemplar of the "Brussels Effect" in action than the DPD itself, which has become a de facto standard for data privacy protection across the globe.[194] Since its enactment in 1995, more than thirty

http://online.wsj.com/article/SB119334720539572002.html [https://perma.cc/KT8A-RNCZ].

[191] Bradford, *supra* note 190, at 3, 18; *see, e.g.*, Daniel W. Drezner, *Globalization, Harmonization, and Competition: The Different Pathways to Policy Convergence*, 12 J. EUROPEAN PUB. POL'Y 841, 841–59 (2005) ("[A] . . . reasonable conjecture would be to say that the public good benefits from regulatory coordination depend upon the size of the newly opened market."); Beth Simmons, *The International Politics of Harmonization: The Case of Capital Market Regulation*, *in* DYNAMICS OF REGULATORY CHANGE: HOW GLOBALIZATION AFFECTS NATIONAL REGULATORY POLICIES 42, 50–52 (2001); David A. Wirth, *The EU's New Impact on U.S. Environmental Regulation*, 31 FLETCHER F. WORLD AFF. 91, 96 (2007) ("If [a] jurisdiction's market share is sufficiently large, [its] regulatory requirements can affect an even larger area, including those under the control of other sovereign authorities.").

> This process can be distinguished from political globalization of regulatory standards where regulatory convergence results from negotiated standards, including international treaties or agreements among states or regulatory authorities. It is also different from unilateral coercion, where one jurisdiction imposes its rules on others through threats or sanctions. Unilateral regulatory globalization is a development where a law of one jurisdiction migrates into another in the absence of the former actively imposing it or the latter willingly adopting it."

Bradford *supra* note 190, at 4.

[192] *See* Bradford, *supra* note 190, at 1.

[193] *See id.* at 3.

[194] *See id.*

countries have heeded Brussels' call by "adopt[ing] EU-type privacy laws, including most countries participating in the Organization for Economic Cooperation and Development."[195]

According to those who have studied the "Brussels Effect" closely, its underlying mechanics are relatively intuitive. Countries confronted with the EU regulations' stringent standards face a stark choice. They can either revise their own domestic policies to reflect those within Europe or risk breaking economic ties with the world's largest trading bloc.[196] For most, the decision requires little more than a moment's contemplation. Aside from a few notable outliers—such as the United States,[197] Russia, and China—most countries simply make the rational calculation that the costs of exclusion from a market consisting of 500 million of the globe's most affluent inhabitants far outweigh the costs of complying with Europe's higher standards.[198]

And lest those powerful incentives prove to be insufficient, the GDPR also includes a number of notable changes intended to promote extraterritorial compliance that are likely to extend its regulatory reach above and beyond the baseline already established by the "Brussels Effect." The most significant changes, in this realm, are those involving the Regulation's "adequacy decision" used to determine whether "third countries" (i.e., countries outside of the EU) have sufficient protections in place to warrant the transfer of personal data between themselves and EU Member States.[199] Once a country is deemed "adequate" through an assessment by the European Commission, data can flow freely without the need for additional protective measures.[200] But unlike the DPD, adequacy decisions made under the GDPR will be subject to a periodic review at least once every four years and will also be subject to

[195] *See id.* at 23.

[196] *See* David Bach & Abraham L. Newman, *The European Regulatory State and Global Public Policy: Micro-Institutions, Macro-Influence*, 14 J. EUROPEAN PUB. POL'Y 827, 831 (2007); Bradford, *supra* note 190, at 11–28. There are, of course, other factors that contribute to this effect. *See id.* at 11–19.

[197] *See* Bradford, *supra* note 190, at 13, 15.

[198] The EU's population exceeds 500 million, and its GDP per capita exceeds $35,000. *See Living in the EU*, EUROPEAN UNION, https://europa.eu/european-union/about-eu/figures/living_en [https://perma.cc/YD47-J682] (last visited Apr. 3, 2019); *European Union GDP Per Capita Ppp*, TRADING ECON., https://tradingeconomics.com/european-union/gdp-per-capita-ppp [https://perma.cc/QU8X-K74N] (last visited Apr. 3, 2019).

[199] *See* GDPR, *supra* note 19, at art. 45.

[200] *See id*; *see also, e.g.*, Press Release, European Comm'n, Questions & Answers on the Japan Adequacy Decision (July 17, 2018) (describing an adequacy decision as "a decision taken by the European Commission establishing that a third country provides a comparable level of protection of personal data to that in the European Union, through its domestic law or its international commitments").

repeal, amendment, or suspension on an ongoing basis.[201]

Thanks to the introduction of these far-reaching forms of regulatory oversight, the GDPR is already showing signs of its global standard-setting authority. Countries such as Israel, New Zealand, Argentina, and Japan have all recently undergone efforts to receive EU "adequacy" certifications by ensuring that their domestic data protections rise to the level of Europe's.[202] "Other countries, from Colombia to South Korea to the tiny island nation of Bermuda, are similarly rebooting [their] domestic legislation . . . [which at times] involves adopting European rules almost word for word."[203]

B. THE GDPR'S EFFECTS ON GLOBAL ENTERPRISE

Though the "Europeanization" of global regulatory standards is often most pronounced at the national level, a phenomenon like the one occurring on the global scale due to the "Brussels Effect" is also taking place within individual enterprises. According to a recent headline-grabbing announcement by Facebook, "[d]ozens of people at [the company] are working full time on" GDPR compliance—requiring upwards of a 250% increase in staffing related to EU data protection.[204] A company spokesperson noted:

> It is hard for us to put an exact figure on it, but when you take into account the time spent by our existing teams, the research and legal assessments and the fact that we have had to pull in teams from product and engineering, it is likely to be millions of dollars.[205]

Recent reporting by *The Financial Times* provided even further confirmation of this phenomenon. The media outlet—which contacted twenty "of the largest social media, software, financial technology and internet companies with EU operations"—noted that its inquiries "revealed that the sector is scrambling to hire new staff and redesign products as it faces millions of dollars in higher costs and lost revenues."[206] And while not every company has quite the multinational reach of the average tech giant, this extraterritorial effect is

[201] *See* GDPR, *supra* note 19, at art. 45.

[202] *See* Mark Scott & Laurens Cerulus, *Europe's New Data Protection Rules Export Privacy Standards Worldwide*, POLITICO (Jan. 31, 2018), https://www.politico.eu/article/europe-data-protection-privacy-standards-gdpr-general-protection-data-regulation/ [https://perma.cc/DRX2-Y9BZ].

[203] *Id.*

[204] Aliya Ram, *Tech Sector Struggles to Prepare for New EU Data Protection Laws*, FIN. TIMES (Aug. 29, 2017), https://www.ft.com/content/5365c1fa-8369-11e7-94e2-c5b903247afd [https://perma.cc/S6GS-RXPW].

[205] *Id.*

[206] *Id.* This phenomenon has led some experts to speculate that the "GDPR could be one of the most expensive pieces of regulation in the [technology] sector's history." *Id.*

made all the more pronounced by the GDPR's applicability to *any* company processing the data of EU citizens, not just those companies actually located within the EU itself.[207]

For some companies operating outside of the GDPR's immediate purview, it may be feasible to fragment their internal processing pipelines by treating data originating in Europe differently from that of other geographies. But doing so could prove administratively onerous and require multiple, separate handling processes for data flowing through any given enterprise. Moreover, this type of maneuver may also be perceived as a public relations risk for companies concerned about being "outed as deliberately offering a lower privacy standard to [their] home users [versus] customers abroad."[208] Thus, just as is true at the national level, the path of least resistance for many companies will likely entail treating the GDPR as the new "gold standard." Ultimately, the Regulation enforcement agencies will effectively dictate the way companies handle all personal data, regardless of geography.[209] While the precise contours of this new gold standard may be continuously revised, it is now clear that it includes a muscular "right to explanation" with sweeping implications for companies and countries throughout the world. As one commentator working to promote GDPR compliance as far away as South Africa recently noted, any entity not currently addressing it will soon realize that the "GDPR has long tentacles."[210]

VI. CONCLUSION

Now that the data protection authorities responsible for enforcing the GDPR's "right to explanation" have weighed in, at least one matter of fierce public debate appears closer to resolution. The GPDR's enforcement

[207] *See* GDPR, *supra* note 19, at art. 3 (describing the territorial scope of the Regulation as applying to any entities "processing . . . personal data of data subjects who are in the Union"); *see also, e.g.*, Goodman & Flaxman, *supra* note 46, at 2 (commenting that the GDPR's "requirements do not just apply to companies that are headquartered in the EU but, rather, to any companies processing EU residents' personal data . . . [thus] [f]or the purposes of determining jurisdiction, it is irrelevant whether that data is processed within the EU territory, or abroad"); Lomas, *supra* note 30 (noting "that GDPR does not merely apply to EU businesses; any entities processing the personal data of EU citizens need to comply").

[208] *See* Lomas, *supra* note 30.

[209] *See* GDPR, *supra* note 19, at art. 3 (describing the territorial scope of the Regulation as applying to any entities "processing . . . personal data of data subjects who are in the Union"); *The History of the General Data Protection Regulation*, EUROPEAN DATA PROTECTION SUPERVISOR, https://edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en [https://perma.cc/2SZ9-Y4ZP].

[210] Scott & Cerulus, *supra* note 202.

authorities envision a muscular "right to explanation" with sweeping legal implications for the design, prototyping, field testing, and deployment of automated data processing systems. Failing to countenance this right could subject enterprises to economic sanctions of truly historic magnitudes—a threat that simply did not exist under the GDPR's predecessor.

Although the protections enshrined by the right may not mandate transparency in the form of a complete individualized explanation, a holistic examination of the Regulation reveals that the right's true power derives from its synergies with other DPbD practices codified by the Regulation's subsequent chapters. While these new design standards will undoubtedly pose significant challenges for the enterprises that fall within the GDPR's purview, the speed and scale of the global response thus far are cause for genuine optimism. Indeed, there is perhaps no more hopeful bookend to this profoundly important debate than the recent words of Bryce Goodman, one of the authors responsible for first sparking the controversy: "In the past, companies have devoted immense resources to improving algorithmic performance. Going forward, one hopes to see similar investments in promoting fair and accountable algorithms."[211]

[211] Bryce Goodman, *supra* note 186, at 7.

THE RIGHT TO EXPLANATION, EXPLAINED

Margot E. Kaminski[†]

ABSTRACT

Many have called for algorithmic accountability: laws governing decision-making by complex algorithms, or artificial intelligence (AI). The EU's General Data Protection Regulation (GDPR) now establishes exactly this. The recent debate over the "right to explanation" (a right to information about individual decisions made by algorithms) has obscured the significant algorithmic accountability regime established by the GDPR. The GDPR's provisions on algorithmic accountability, which include a right to explanation, have the potential to be broader, stronger, and deeper than the requirements of the preceding Data Protection Directive. This Article clarifies, including for a U.S. audience, what the GDPR requires.

DOI: https://doi.org/10.15779/Z38TD9N83H

[†] Associate Professor of Law, University of Colorado Law School. Faculty Director of Privacy at Silicon Flatirons. Affiliated Fellow, Information Society Project at Yale Law School. Thanks to Andrea Bertolini, Kiel Brennan-Marquez, Giovanni Comandé, Matthew Cushing, Natalie Helberger, Max van Drunen, Nico van Eijk, Sarah Eskens, Gianclaudio Malgieri, Nicholson Price, Marijn Sax, and Andrew Selbst, and to the Fulbright-Schuman program, IViR at University of Amsterdam, and Scuola Superiore Sant'Anna for their support. All errors are my own.

TABLE OF CONTENTS

I. INTRODUCTION

Scholars and civil society groups on both sides of the Atlantic have been calling for algorithmic accountability: laws governing decision-making by complex algorithms, or AI.[1] Algorithms can be used to make, or to greatly

[1] *See, e.g.*, Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC. 973 (2016) (analyzing the benefits and limitations of transparency in establishing algorithmic accountability); Lee A. Bygrave, *Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling*, 17 COMPUTER L. & SECURITY REP. 17 (2001) (analyzing Art. 15 of the 1995 EC Directive on data protection); Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008) (examining algorithmic decision-making and calling for transparency, accountability, and accuracy); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014) (calling for accountability for automated predictions); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014) (charting the privacy harms caused by big data and proposing procedural due process); Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1 (2017) (providing a computer scientist's perspective on algorithmic accountability and calling for specific tailored solutions); Mireille Hildebrandt, *The Dawn of a Critical Transparency Right for the Profiling Era*, DIGITAL ENLIGHTENMENT Y.B. 41 (2012) (highlighting the potential of the GDPR to protect individuals in the profiling era); Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189 (2017) (proposing a toolkit to ensure algorithmic accountability); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017) (calling for collaboration on algorithmic accountability across computer science, law, and policy); W. Nicholson Price II, *Regulating Black-Box Medicine*, 116 MICH. L. REV. 421 (2017) (proposing that black-box medical algorithms should be governed through collaborative governance); Neil M. Richards & Jonathan H. King, *Big Data Ethics*, 49 WAKE FOREST L. REV. 393 (2014) (calling for ethical standards to be applied to mass data collection

affect, decisions about credit, employment, education, and more.[2] Algorithmic decision-making can be opaque, complex, and subject to error, bias, discrimination, in addition to implicating dignitary concerns.[3] The literature in

and use); Paul Schwartz, *Data Processing and Government Administration: The Failure of the American Legal Response to the Compute*r, 43 HASTINGS L.J. 1321 (1992) (developing an approach to govern the use of computers and personal data); Daniel J. Steinbock, *Data Matching, Data Mining, and Due Process*, 40 GA. L. REV. 1, 23 (2005) (addressing the use of data matching and mining to identify persons against whom an official action is taken); Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83 (2017) (calling for a federal agency to govern algorithms); Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503 (2013) (creating a framework for understanding transparency as a regulatory concept in algorithmic accountability); Michal S. Gal, *Algorithms as Illegal Agreements*, 34 BERKELEY TECH. L.J. 67 (2019) (examining potential legal solutions to concerns raised by algorithmic-facilitated coordination); Bryan Casey, Ashkon Farhangi & Roland Vogl, *Rethinking Explanable Machines: the GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L.J. 145 (2019) (discussing machine explainability in the context of the European GDPR's "right to explanation").

[2] *See, e.g.*, Citron & Pasquale, *supra* note 1, at 4.

[3] *See generally* Margot E. Kaminski, *Binary Governance*, 92 S. CAL. L. REV. (forthcoming Sept. 2019) (identifying three categories of concerns behind calls for regulating algorithmic decision-making: dignitary, justificatory, and instrumental); *see also* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1118–26 (2018) (discussing the rationales behind calls for explanations of algorithmic decision-making) [hereinafter Selbst & Barocas, *Intuitive Appeal*]. On error, see Citron & Pasquale, *supra* note 1, at 8 ("Scoring systems and the arbitrary and inaccurate outcomes they produce must be subject to expert review."); Crawford & Schultz, *supra* note 1, at 104 ("This aggregation of various agencies' data allows law enforcement to predict or flag individuals as suspicious or worthy of investigation, search, or detention based on the agency's outlined criteria . . . . [T]his method may sometimes lead to erroneous results."); Zarsky, *supra* note 1, at 1506 (noting that "the growing use of predictive practices . . . could be tainted with errors and overinvasive"). On bias and discrimination, see Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 674 (2016) ("Approached without care, data mining can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society."); Citron, *supra* note 1, at 1262 (noting that "[t]he biases of individual programmers can have a larger, accumulating effect"); Citron & Pasquale, *supra* note 1, at 13 ("Far from eliminating existing discriminatory practices, credit-scoring algorithms instead grant them an imprimatur, systematizing them in hidden ways."). On dignity, see Bygrave, *supra* note 1, at 18; Isak Mendoza & Lee A. Bygrave, *The Right Not to Be Subject to Automated Decisions Based on Profiling*, *in* EU INTERNET LAW: REGULATION AND ENFORCEMENT 77, 84 (Tatiani Synodinou et al. eds., Springer, 2017) (noting "a concern to uphold human dignity by ensuring that humans (and not their 'data shadows') maintain the primary role in 'constituting' themselves"); Zarsky, *supra* note 1, at 1548; *see also* Meg Leta Jones, *The Right to A Human in the Loop: Political Constructions of Computer Automation and Personhood*, 42 SOC. STUD. SCI. 216 (2017) (exploring the role of dignity in data protection law addressing automated decision-making).

the United States has been largely speculative, operating in a policy vacuum.[4] This is resolutely not, however, the case in the European Union.

On May 25, 2018, the General Data Protection Regulation (GDPR) went into effect in the EU.[5] The GDPR contains a significant set of rules on algorithmic accountability, imposing transparency, process, and oversight on the use of computer algorithms to make significant decisions about human beings.[6] The GDPR may prove to be an example, both good and bad, of a robust algorithmic accountability regime in practice.[7] However, to a U.S. audience, the recent vigorous debate around whether there is a "right to explanation" in the GDPR may inspire confusion.[8] Arguments over the

[4] Senator Wyden has, for example, proposed algorithmic accountability as part of his proposed federal privacy legislation. More recently, Senator Wyden and Senator Booker along with Representative Clarke proposed the Algorithmic Accountability Act of 2019. Federal law governing the private sector's use of algorithmic decision-making does not, however, currently exist. *See* S. 2188, 115th Cong. (2018), at 2, 6, 32; *see also* S. _ 116th Cong. (2019) (Algorithmic Accountability Act of 2019).

[5] *GDPR FAQs*, EU GDPR.ORG, https://eugdpr.org/the-regulation/gdpr-faqs/ [https://perma.cc/FV79-VBRU] (last visited Mar. 13, 2019).

[6] Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 at arts. 22, 13, 14, 15 [hereinafter GDPR].

[7] *Compare, e.g.*, Tal Z. Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 SETON HALL L. REV. 995, 1014–15 (2017), *with* Hildebrandt, *supra* note 1.

[8] *See* Maja Brkan, *Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond*, INT'L J.L. & INFO. TECH. 1, 13–20 (2019); Casey et al., *supra* note 1; Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking for*, 16 DUKE L. & TECH. REV. 17, 44 (2017) [hereinafter Edwards & Veale, *Slave to the Algorithm*]

> In 2016, to the surprise of some EU data protection lawyers, and to considerable global attention, Goodman and Flaxman asserted in a short paper that the GDPR contained a "right to an explanation" of algorithmic decision making. As Wachter et al. have comprehensively pointed out, the truth is not quite that simple.

Lilian Edwards & Michael Veale, *Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?*, 16 IEEE SECURITY & PRIVACY 46 (2018); Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision-Making and "a Right to Explanation"*, 38 AI MAG. 50, 55–56 (2017); Gianclaudio Malgieri & Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 243, 246 (2017); Mendoza & Bygrave, *supra* note 3, at 16; Antoni Roig, *Safeguards for the Right Not to Be Subject to a Decision Based Solely on Automated Processing (Article 22 GDPR)*, 8 EURO. J. L. & TECH. 1 (2017); Selbst & Barocas, *Intuitive Appeal*, *supra* note 3, at 1106; Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT'L DATA PRIVACY L. 233, 235 (2017); Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76 (2017) [hereinafter Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does not Exist*]; Sandra Wachter, Brent Mittelstadt & Chris Russell, purported right to explanation obscure the true substance and depth of the GDPR's algorithmic accountability regime.

This Article clarifies, including for a U.S. audience, what is and is not required by the GDPR. It contributes to the existing conversation over algorithmic accountability in the GDPR by addressing the authoritative guidelines on automated decision-making.[9] Contrary to several scholars, I understand the GDPR to create a broader, stronger, and deeper algorithmic accountability regime than what existed under the EU's Data Protection Directive (DPD).[10] The debate over the right to explanation threatens to obscure this significant development.

Part II of this Article begins by explaining for a U.S. audience the status of the various interpretative documents that accompany the GDPR. Part III identifies the provisions of the GDPR that apply to algorithmic accountability, and points to textual ambiguities that gave rise to disagreements over the right to explanation. Part IV uses the interpretative documents introduced in Part II, including recent authoritative guidelines, to show how many of the questions left open in the GDPR's text have been subsequently narrowed or resolved. Part V turns to the right to explanation and other transparency mechanisms. Throughout, this Article focuses on the GDPR's requirements for private companies rather than for governments.

II. GDPR BASICS

First, a U.S. audience needs to understand the legal materials at play. The GDPR consists of both text (Articles) and an extensive explanatory preamble. The preambular provisions, known as Recitals, do not have the direct force of law in the EU.[11] A Recital is supposed to "cast light on the interpretation to be

Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, 31 HARV. J. L. & TECH. 841 (2018) [hereinafter Wachter et al., *Counterfactual*].

[9] ARTICLE 29 DATA PROTECTION WORKING PARTY, GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING AND PROFILING FOR THE PURPOSES OF REGULATION 2016/679, 17/EN. WP 251rev.01 (Feb. 6, 2018) [hereinafter GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING]. Three pieces address the earlier draft version of these guidelines. Casey et. al, *supra* note 1, at 171; *see generally* Michael Veale & Lilian Edwards, *Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling*, 2 COMPUT. L. & SECURITY REV. 398 (NEEDS PARA); Wachter et. al, *Counterfactual*, *supra* note 8.

[10] *See* Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 20–21; Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist*, *supra* note 8, at 78; Wachter et. al, *Counterfactual*, *supra* note 8, at 861–71.

[11] Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does not Exist*, *supra* note 8, at 80.

given to a legal rule [but] it cannot in itself constitute such a rule."[12] This gives Recitals a liminal legal status—they are not binding law, but they are often cited as authoritative interpretations where the GDPR is vague.[13]

While Recitals can clarify how the GDPR's standards should be applied, they often contain language that goes well beyond what is in the GDPR itself, reflecting the result of political compromise during negotiations.[14] Recitals cannot create new legal requirements, but the line between valid interpretation and invalid creation of new law can be hard to draw.

Discussions of the GDPR also frequently cite interpretative guidelines issued by a group previously known as the Article 29 Working Party and now called the European Data Protection Board.[15] The Working Party/Data Protection Board is made up of Data Protection Authorities (the regulators tasked with enforcing the GDPR) from around the EU who come to a consensus over the interpretation of data protection provisions. As Data Protection Authorities in EU Member States enforce the GDPR on the ground, they refer to the guidelines issued by the Working Party/Data Protection Board.

Article 29 Working Party guidelines, again, do not have the direct force of law. They are, nonetheless, strongly indicative of how enforcers will interpret the law. Now that the GDPR is in effect, these guidelines have additional, though indirect, teeth. The European Data Protection Board under the GDPR has additional supervisory and harmonizing capabilities over Member State Data Protection Authorities.[16] A local Data Protection Authority, in other words, is now even more likely to adhere to the guidelines than under the Directive.

U.S. audiences thus need to understand that while only the text of the GDPR is technically binding law, both Recitals and Working Party/Data Protection Board guidelines play a significant role, in practice, in guiding how

[12] Case 215/88 Casa Fleischhandels [1989] European Court of Justice ECR 2789 [31], https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A61988CJ0215 [https://perma.cc/C8RK-45FP].

[13] *See, e.g.*, Brkan, *supra* note 8, at 16 ("Dismissing the possibility of the existence of the right to explanation altogether because recitals are not legally binding is too formalistic.").

[14] Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 50 ("In the GDPR however, as a matter of political expediency, many issues too controversial for agreement in the main text have been kicked into the long grass of the recitals, throwing up problems of just how binding they are.").

[15] GDPR, *supra* note 6, at art. 68.

[16] *Id.* at art. 70; *see* Amber Hawk, *The Recitals Are Essential to Your Understanding the General Data Protection Regulation*, HAWK TALK (Jan. 28, 2016), http://amberhawk.typepad.com/amberhawk/2016/01/the-recitals-are-essential-to-your-understanding-the-general-data-protection-regulation.html [https://perma.cc/7S5B-AH8E].

companies will behave. A company, concerned about the GDPR's significant penalties (famously up to 4% of worldwide revenue) backed by an increasingly rights-protective European Court of Justice, is likely to follow both the Recitals and Working Party guidance because they are indicative of what the GDPR's enforcers are likely to do.[17] Although these texts are not technically binding, they strongly indicate how enforcers and eventually courts will likely interpret the text.

In another Article, I argue at length that this is precisely how the GDPR is intended to work.[18] The GDPR is, in large part, a collaborative governance regime.[19] The text is full of broad standards, to be given specific substance over time through ongoing dialogues between regulators and companies, backed eventually by courts. Both the Recitals and the Working Party guidelines, along with numerous mechanisms ranging from a formal process for establishing codes of conduct to less formal impact assessment requirements, are part of this collaborative approach.[20]

Thus, when scholars argue that what is in the Recitals is not the law,[21] they are not only insisting on a technicality—distinguishing between harder and softer legal instruments—they are also disregarding the fundamentally collaborative, evolving nature of the GDPR, and removing important sources of clarity for companies as the law develops.

[17] GDPR, *supra* note 6, at art. 84. For indicators of the Court's increasing interest in data protection, see, for example, Joint Cases C-293/12 & C-594/12, Digital Rights Ireland Ltd. v. Minister for Commc'ns, ECLI:EU:C:2014:238 (2014) (finding data retention requirements to violate the fundamental right to data protection); Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (AEPD) (2014) (finding that Google as a search engine is a data controller and thus is responsible for affording individuals the data protection right to erasure ("right to be forgotten") from search engine indexing). While the ECtHR is not responsible for GDPR interpretation, it also forms a backstop to surveillance-related law in the EU. *See* Roman Zakharov v. Russia, 2015-VIII Eur. Ct. H.R. 205 (finding Russian metadata surveillance in violation of fundamental rights).

[18] *See* Kaminski, *Binary Governance*, *supra* note 3.

[19] For discussions of collaborative governance (also known as "new governance"), see, e.g., Jody Freeman, *Collaborative Governance in the Administrative State*, 45 UCLA L. REV. 1 (1997); Orly Lobel, *The Renew Deal: The Fall of Regulation and the Rise of Governance in Contemporary Legal Thought*, 89 MINN. L. REV. 342 (2004).

[20] Kaminski, *Binary Governance*, *supra* note 3, at 21–22.

[21] *See* Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist*, *supra* note 8, at 80.

III. ALGORITHMIC ACCOUNTABILITY IN THE TEXT OF THE GDPR

This Part introduces the text of the GDPR that applies to algorithmic decision-making. There are four Articles of the GDPR that specifically address algorithmic decision-making. Article 22 of the GDPR addresses "[a]utomated individual decision-making, including profiling."[22] Articles 13, 14, and 15 each contain transparency rights around automated decision-making and profiling.[23] More general GDPR provisions, such as the right to object, the right to rectification (correction), data protection by design and by default, and the requirement of data protection impact assessments, likely apply to most or even all algorithmic decision-making.[24] For the sake of brevity and clarity, this Part discusses only the text of Articles 22, 13, 14, and 15, which specifically reference automated decision-making.[25] As others have pointed out, however, the more generally applicable provisions of the GDPR also play an important role in governing algorithmic decision-making.[26]

A. ARTICLE 22: AUTOMATED INDIVIDUAL DECISION-MAKING

Article 22 states that individuals "have the right not to be subject to a decision based solely on automated processing."[27] Scholars have pointed out, based on the historical treatment of similar text in the Data Protection Directive (DPD), the predecessor to the GDPR, that this could be interpreted as either a right to object to such decisions or a general prohibition on significant algorithmic decision-making. [28] Interpreting Article 22 as establishing a right to object would make the right narrower. In practice, it would allow companies to regularly use algorithms in significant decision-making, adjusting their behavior only if individuals invoke their rights.

[22] GDPR, *supra* note 6, at art. 22.

[23] *See id.* at arts. 13(2)(f), 14(2)(g), 15(1)(h).

[24] *See* Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 19 (noting "other parts of the GDPR related (i) to the right to erasure ('right to be forgotten') and the right to data portability; and (ii) to privacy by design, Data Protection Impact Assessments and certification and privacy seals"), 23, 77; Casey et. al, *supra* note 8, at 173–76 (discussing DPIA safeguards); GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 29 (discussing DPIA and data protection officer), 34 (discussing right to object); *see also* GDPR, *supra* note 6, Recital 91 (described as "[n]ecessity of a data protection impact assessment").

[25] *See generally* GDPR, *supra* note 6, at arts. 13, 14, 15, 22.

[26] *See* Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 19.

[27] GDPR, *supra* note 6, at art. 22(1).

[28] *See, e.g.*, Mendoza & Bygrave, *supra* note 3, at 9 ("[t]his distinction . . . suggests that Art. 22(1) is intended as a prohibition and not a right that the data subject has to exploit" but noting that it can be argued both ways); Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does not Exist*, *supra* note 8, at 94.

Interpreting Article 22 instead as a prohibition on algorithmic decision-making would require all companies using algorithmic decision-making to assess which exception they fall under and to implement safeguards to protect individual rights, or to not deploy algorithmic decision-making at all.

The Article 22 right/prohibition applies only when the decision is "based solely" on algorithmic decision-making, and it applies only when the decision produces "legal effects" or "similarly significant" effects on the individual.[29] What either of these restrictions means is unclear from the GDPR's text alone. One could narrowly interpret "based solely" to mean that any human involvement, even rubber-stamping, takes an algorithmic decision out of Article 22's scope; or one could take a broader reading to cover all algorithmically-based decisions that occur without *meaningful* human involvement.[30] Similarly, one could take a narrow reading of "similarly significant" effects to leave out, for example, behavioral advertising and price discrimination, or one could take a broader reading and include behavioral inferences and their use.[31]

There are three exceptions to the Article 22 right/prohibition. The first is when the automated decision is "necessary for . . . a contract."[32] The second is when a Member State of the European Union has passed a law creating an exception.[33] The third is when an individual has explicitly consented to algorithmic decision-making.[34] Both the contractual exception and the explicit consent exception could be interpreted to be broader or narrower in nature,

[29] GDPR, *supra* note 6, at art. 22(1) ("The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.").

[30] *See, e.g.*, Mendoza & Bygrave, *supra* note 3, at 11 ("Even if a decision is formally ascribed to a person, it is to be regarded as based solely on automated processing if a person does not actively assess the result of the processing prior to its formalization as a decision."); Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does not Exist*, *supra* note 8, at 88 ("[T]his creates a loophole whereby even nominal involvement of a human in the decision-making process allows for an otherwise automated mechanism to avoid invoking elements of the right of access . . . addressing automated decisions.").

[31] *See, e.g.*, Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 47–48 (discussing whether advertising constitutes a significant effect), 69 (discussing the GDPR's inconsistent treatment of inferences); Malgieri & Comandé, *supra* note 8, at 265 ("[S]ignificant effects should also include cases of neuromarketing manipulation or price discrimination . . . .").

[32] GDPR, *supra* note 6, at art. 22(2)(a) ("[N]ecessary for entering into, or performance of, a contract between the data subject and a data controller.").

[33] *Id.* at art. 22(2)(b) ("[A]uthorised by Union or Member State law to which the controller is subject.").

[34] *Id.* at art. 22(2)(c) ("[B]ased on the data subject's explicit consent.").

depending for example on how one interprets "necessary for . . . a contract."[35] In the case of sensitive, or "special category," data, even fewer exceptions apply.[36]

Even when an exception to Article 22 applies, a company must implement "suitable measures to safeguard the data subject's rights and freedoms and legitimate interests . . . ."[37] This requirement is the source of the debate over the right to explanation. Suitable safeguards, according to the text, must include "at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision."[38] This explicitly creates a version of algorithmic due process: a right to an opportunity to be heard.[39] These are the only safeguards named in the GDPR's text. The use of the words "at least," however, indicates that these are an open list of minimum requirements, and a company should do more. As discussed in Part IV, both the preamble (Recital) and interpretative guidance have added to this list of both suggested and required safeguards, and both include as a safeguard a right to explanation of an individual decision.

One important note on suitable safeguards: the specific minimum examples above apply with respect to the contractual exception and explicit consent exception, but are not in the text of the Member State law exception.[40] This textual difference leaves room for the possibility that Member States might enact a different set of suitable safeguards.[41] It remains to be seen whether Data Protection Authorities and courts will allow Member States to adopt significantly different protections against algorithmic decision-making.

[35] Mendoza & Bygrave, *supra* note 3, at 14–15; Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does not Exist*, *supra* note 8, at 98.

[36] GDPR, *supra* note 6, at art. 22(4) ("Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies . . . .").

[37] GDPR, *supra* note 6, at arts. 22(2)(b), 22(3).

[38] GDPR, *supra* note 6, at art. 22(3).

[39] Several U.S. scholars have called for algorithmic due process, mimicking procedural due process rights. *See* Citron & Pasquale, *supra* note 1; *see generally* Crawford & Schultz, *supra* note 1.

[40] GDPR, *supra* note 6, at arts. 22(2)(b), 22(3).

[41] Wachter et al. argue that this means that the same safeguards do not apply. Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does not Exist*, *supra* note 8, at 93; Brkan, *supra* note 8, at 12 (describing German law); *see* Gianclaudio Malgieri, Automated Decision-Making in the EU Member States; The Right to Explanation and other 'Suitable Safeguards' (Aug. 17, 2018) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3233611 [https://perma.cc/WLC6-X8QC].

B. Articles 13, 14, and 15: Notification and Access Rights

Outside of Article 22, the GDPR contains a series of individual notification and access rights specific to automated decision-making. Article 13 establishes a series of notification rights/requirements when information is collected directly from individuals.[42] Article 14 establishes a similar set of notification rights/requirements when information about individuals is collected from third parties.[43] Article 15 creates an individual right of access to information held by a company that can be invoked "at reasonable intervals."[44] All three Articles contain an identical provision requiring disclosure of "the *existence* of automated decision-making, including profiling."[45] Additionally, this provision requires disclosure of "*meaningful information about the logic involved*, as well as the significance and the envisaged consequences of such processing for the data subject."[46]

This language has provoked debate, especially over the question of timing.[47] The language in all three Articles is identical, but the temporal context is different. Articles 13 and 14, roughly speaking, require companies to notify individuals when data is obtained,[48] while Article 15 creates access rights at almost any time. Some scholars have argued that because the text of the three Articles is identical, it must refer to the same information, which indicates that "meaningful information about the logic involved" can be only a broad

[42] GDPR, *supra* note 6, at art. 13.

[43] *Id.* at art. 14.

[44] *Id.* at art. 15. *See* GDPR, *supra* note 6, Recital 63 (described as "[r]ight of access").

[45] GDPR, *supra* note 6, at arts. 13(2)(f), 14(2)(g), 15(1)(h) (collectively, "meaningful information" provisions) (emphasis added).

[46] GDPR, *supra* note 6, at arts. 13(2)(f), 14(2)(g), 15(1)(h).

[47] *See, e.g.*, Mendoza & Bygrave, *supra* note 3, at 16 ("[T]he wording of Art. 15 does not *necessarily* exclude the possibility that it embraces a right of ex post explanation of an Art. 22 type decision."); Selbst & Powles, *supra* note 8, at 236; Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does not Exist*, *supra* note 8, at 90 ("As the scope of information data controllers are required to disclose in Article 15 is the same as in Article 13, Article 15 similarly requires only limited information about the functionality of the automated decision-making system.").

[48] GDPR, *supra* note 6, at art. 13 (requiring it when data is obtained); *id.* at art. 14(3)(a) (requiring disclosure "within a reasonable period after obtaining the personal data, but at the latest within one month, having regard to the specific circumstances in which the personal data are processed"). Article 14 also envisions notification in communication with a data subject where data is used for communication (art. 14(3)(b)) or upon disclosure of data to another third party (art. 14(3)(c)). These both refer to a later notification than upon obtaining data, but it is harder to envision when this might refer to algorithmic decision-making that has already occurred (unless one is communicating the results to an individual or third party, perhaps?).

overview of a decision-making system.[49] Others argue, however, that, read in context, "meaningful information" must mean multiple things.[50] Articles 13 and 14 might require an overview of a system prior to processing, but Article 15's access right could provide deeper disclosure, including insight into a particular decision affecting a particular individual. The text of the GDPR does not clarify this conflict one way or another.

There are exceptions to the GDPR's notification and access requirements.[51] While not included in the text of the GDPR, an accompanying Recital mentions an exception for intellectual property rights—that is, trade secrets and copyright law.[52] Some scholars argue that, in practice, trade secrets, in particular, represent a significant obstacle to meaningful disclosure of algorithms.[53] This has certainly been the case in the United States.[54] Others observe, however, that fundamental rights such as the right to data protection take precedence over trade secrecy.[55]

The text of the GDPR thus creates both transparency and process rights around algorithmic decision-making. The text itself, however, leaves considerable room for interpretation. But both accompanying and subsequent interpretative documents narrow and clarify the GDPR's text, resolving a number of the conflicts discussed above.

[49] Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist*, *supra* note 8, at 82.

[50] *See, e.g.*, Malgieri & Comandé, *supra* note 8, at 244; Mendoza & Bygrave, *supra* note 3, at 16; Selbst & Powles, *supra* note 8, at 236.

[51] *See* GDPR, *supra* note 6, at arts. 14(5), 15(4).

[52] *See* GDPR, *supra* note 6, Recital 63 ("That right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software."). The copyright argument makes little sense. *See* Brkan, *supra* note 8, at 22.

[53] Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist*, *supra* note 8, at 85.

[54] *See, e.g.*, Jessica M. Eaglin, *Constructing Recidivism*, 67 EMORY L.J. 59, 111 (2017) ("Transparency Measures"); David S. Levine, *The Impact of Trade Secrecy on Public Transparency*, *in* THE LAW AND THEORY OF TRADE SECRECY: A HANDBOOK OF CONTEMPORARY RESEARCH 406 (Rochelle C. Dreyfuss & Katherine J. Strandburg eds., 2010); Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1349–50 (2018).

[55] Brkan, *supra* note 8, at 21–24; Malgieri & Comandé, *supra* note 8, at 262; Selbst & Powles, *supra* note 8, at 242.

IV. ALGORITHMIC ACCOUNTABILITY IN THE GDPR, INTERPRETED

Both the Recitals and recently adopted Working Party guidelines clarify the GDPR's text in important ways. Article 22 and the "meaningful information" provisions are not devoid of substance; they create an algorithmic accountability regime that is broader, stronger, and deeper than what existed in Europe prior to the GDPR.[56] This Part first explains how the GDPR's text has been clarified, with reference to the debates discussed in Part III above.[57] It then explains why the GDPR's version of algorithmic accountability is broader, stronger, and deeper than Article 15 of the DPD.

First, the Working Party guidelines clarify that Article 22 is a prohibition on algorithmic decision-making, not a mere right to object to it.[58] This is significant because it clarifies that companies have a duty *not to use* solely automated decision-making, rather than a mere duty to respond to individuals who object to it. Companies using algorithmic decision-making will, therefore, have to assess which exception they fall under (contract, explicit consent, or Member State law), which will often trigger additional disclosures to individuals as companies attempt to obtain explicit consent or to justify why such decision-making is necessary to a contract.[59]

Second, the guidelines explain that for an automated decision to fall outside of Article 22, human involvement must be meaningful.[60] A company does not escape Article 22 solely by having a human rubber-stamp algorithmic decisions.[61] Human oversight must be "carried out by someone who has the authority and competence to change the decision."[62] That person must additionally have access to information beyond just the algorithm's outputs.[63] The GDPR will thus have the effect of requiring companies to think about

[56] *See* GDPR, *supra* note 6, at arts. 13(2)(f), 14(2)(g), 15(1)(h).

[57] For another (more pessimistic) take on the guidelines, see Veale & Edwards, *supra* note 9.

[58] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 19.

[59] *Id.* at 13 ("Controllers seeking to rely upon consent as a basis for profiling will need to show that data subjects understand exactly what they are consenting to . . . .").

[60] *Id.* at 21 ("The controller cannot avoid the Article 22 provisions by fabricating human involvement [, and] must ensure that any oversight of the decision is meaningful, rather than just a token gesture.").

[61] *Id.* ("[I]f someone routinely applies automatically generated profiles to individuals without any actual influence on the result, this would still be a decision based solely on automated processing.").

[62] *Id.*

[63] *See id.* (noting that the controller "should consider all the relevant data" during analysis of the decision).

how they structure their "human in the loop" of algorithmic decision-making to escape Article 22's prohibition or forego its safeguard requirements.[64]

Third, both Recital 71 and the guidelines provide examples of decisions with significant effects. Recital 71 provides examples of credit determinations and e-recruiting practices.[65] The Working Party guidelines explain that "only serious impactful effects" will trigger Article 22.[66] The guidelines provide both a framework for determining what constitutes a significant effect[67] and a list of examples: decisions that affect financial circumstances or access to health services or access to education, or decisions that deny employment or put someone "at a serious disadvantage."[68]

The guidelines additionally, and perhaps surprisingly, explain that some behavioral advertising will be covered.[69] Particularly intrusive advertising targeted at particularly vulnerable data subjects in particularly manipulative ways will trigger Article 22.[70] Differential pricing—showing people different prices based on personal profiles—could also trigger Article 22 if "prohibitively high prices effectively bar someone from certain goods or services."[71] Thus Article 22's algorithmic accountability provisions will reach

[64] *See* Citron & Pasquale, *supra* note 1, at 6–7; *see also* Meg Leta Jones, *The Ironies of Automation Law: Tying Policy Knots with Fair Automation Practices Principles*, 18 VAND. J. ENT. & TECH. L. 77 (2015).

[65] GDPR, *supra* note 6, Recital 71

> [S]uch as automatic refusal of an online credit application or e-recruiting practices without any human intervention . . . in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her.

[66] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 21. Examples of legal effects in the guidelines largely involve government use of algorithms rather than use by private companies but include the cancellation of a contract. The guidelines list as examples: entitlement to or denial of a social benefit granted by law; and immigration effects. *See* ARTICLE 29 DATA PROTECTION WORKING PARTY, GUIDELINES FOR IDENTIFYING A CONTROLLER OR PROCESSOR'S LEAD SUPERVISORY AUTHORITY, 16/EN, WP 244 (Dec. 13, 2016), at 4 (discussing "substantially affects").

[67] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 21 ("[S]ignificantly affect the circumstances, behaviour or choices of the individuals concerned; have a prolonged or permanent impact on the data subject; or at its most extreme, lead to the exclusion or discrimination of individuals.").

[68] *Id.* at 22.

[69] *Id.* ("Similarly significant effects could also be triggered by the actions of individuals other than the one to which the automated decision relates.").

[70] *Id.*

[71] *Id.*

at least some behavioral advertising and some differential pricing tactics. This coverage is broader than some scholars predicted.[72]

Fourth, the Working Party guidelines somewhat close the trade secrets loophole to algorithmic transparency. Several scholars feared that in practice, companies could avoid the GDPR's transparency requirements by citing a need for corporate secrecy.[73] The guidelines explain, however, that while there is "some protection" against having to reveal trade secrets, companies "cannot rely on the protection of their trade secrets as an excuse to deny access or refuse to provide information . . . ."[74] While this does not eliminate the trade secrets exception discussed in Recital 63, it does at least urge data protection authorities to watch for the use of overly broad trade secrets claims.

Fifth, the guidelines clarify that both the contractual exception and the explicit consent exception to Article 22 are relatively narrow.[75] For example, online retailers cannot argue that profiling is necessary for an online purchase, even where profiling is mentioned in the fine print of the contract.[76] Automated decision-making might be necessary where human involvement is impossible due to the sheer quantity of information processed, but then the company must show that there is no other effective and less privacy-intrusive way to accomplish the same goal.[77]

The guidelines similarly constrain the explicit consent exception and turn it into an information-driving tool.[78] They explain that individuals must be provided enough information about the use and consequences of profiling to ensure that any consent "represents an informed choice."[79] The guidelines do not provide additional information about "explicit consent," except to note that while explicit consent is not defined in the GDPR, a "high level of individual control over personal data is . . . deemed appropriate."[80]

[72] *See* Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist*, *supra* note 8, at 92–93, 98; *see, e.g.*, Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 47–48 (questioning whether race-targeted advertising constitutes a significant effect on an individual).

[73] *See* Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does not Exist*, *supra* note 8, at 85–86; *see also* Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 53 ("Article 15(h) has a carve out in the recitals, for the protection of trade secrets and IP.").

[74] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 17.

[75] *Id.* at 13 ("[N]ecessity should be interpreted narrowly.").

[76] *See id.*

[77] *Id.* at 23.

[78] *See id.* at 13, 23.

[79] *Id.* at 13.

[80] *Id.* at 24. *See* ARTICLE 29 DATA PROTECTION WORKING PARTY, GUIDELINES ON CONSENT UNDER REGULATION 2016/679, 17/EN, WP259, (Nov. 28, 2017).

Finally, the guidelines address the central question of what is required as "appropriate safeguards" to protect individuals from automated decision-making when one of the exceptions applies.[81] Scholars have argued that there is no right to an explanation of individual decisions in the GDPR because that right is not specifically enumerated in the GDPR's text.[82] That reasoning is wrong.[83] Recital 71 states that "suitable safeguards . . . should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to *obtain an explanation of the decision reached* after such assessment and to challenge the decision."[84]

The Working Party guidelines directly quote this language, not once but thrice.[85] The guidelines counsel that there is a need for this form of transparency because an individual can challenge a particular decision or express her view only if she actually understands "how it has been made and on what basis."[86] In other words, an individual has a right to explanation of an individual decision because that explanation is necessary for her to invoke the other rights—e.g., to contest a decision, to express her view—that are explicitly enumerated in the text of the GDPR.[87]

Beyond the right to explanation, the guidelines explain that the GDPR establishes a version of individual algorithmic due process by creating an

[81] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 27.

[82] Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 50 ("Our view is that these certainly seem shaky foundations on which to build a harmoni[z]ed cross-EU right to algorithmic explanation."); Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist*, *supra* note 8, at 79.

[83] At this point, the bulk of the literature on the right to explanation appears to agree that this reasoning is erroneous. *See* Brkan, *supra* note 8, at 16 ("Dismissing the possibility of the existence of the right to explanation altogether because recitals are not legally binding is too formalistic, in particular in the light of the CoJ's case law which regularly uses recitals as an interpretative aid."); Malgieri & Comandé, *supra* note 8, at 255 ("[T]he right to obtain an explanation of the decision reached after the assessment should always be exercisable."); Mendoza & Bygrave, *supra* note 3, at 16 ("[W]e should not discount the possibility that a right of ex post explanation of automated decisions is implicit in the right 'to contest' a decision pursuant to Art. 22(3)."); Selbst & Powles, *supra* note 8, at 235 ("Recital 71 is not meaningless, and has a clear role in assisting interpretation and co-determining positive law."), 242 ("We believe that the right to explanation should be interpreted functionally, flexibly, and should, at a minimum, enable a data subject to exercise his or her rights under the GDPR and human rights law.").

[84] GDPR, *supra* note 6, Recital 71 (emphasis added).

[85] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 19, 27, 35.

[86] *Id.* at 27.

[87] Both Mendoza & Bygrave and Selbst & Powles suggested precisely this. Mendoza & Bygrave, *supra* note 3, at 16; Selbst & Powles, *supra* note 8, at 242.

opportunity to be heard.[88] The guidelines note that safeguards must include human intervention by a reviewer with "the appropriate authority and capability to change the decision," and who should have access to "all the relevant data."[89] This imposes another form of transparency, albeit internal to a company, as technical information flows to the human called on to intervene in an algorithmic decision. There is little in the guidelines, however, outlining how human intervention and contestation should take place, apart from suggesting that companies provide a link to an appeals process, a timeline for review, and a named contact person for inquiries.[90] This opportunity to be heard thus may prove to be more or less meaningful, in practice, and risks being, as currently described, reduced to the provision of a contact email.

The next interpretative move that the guidelines make might not be intuitive to a U.S. audience expecting a system entirely focused on individual rights. Beyond individual due process, the guidelines interpret "suitable safeguards" to also include systemic accountability measures such as auditing and ethical review boards.[91] These systemic accountability measures have dual meaning: They can be understood as bolstering individual rights by ensuring that somebody impartial is providing oversight in the name of individuals, or as providing necessary accountability over company behavior in a collaborative governance (private/public partnership) regime, as companies come up with and implement systems for preventing error, bias, and discrimination.[92]

In practice, this systemic accountability involves a number of system-wide checks. Scholars have read Recital 71's language to require algorithmic auditing.[93] The Working Party Guidelines support this interpretation, suggesting that safeguards include quality assurance checks, algorithmic auditing, independent third-party auditing, and more.[94] Both Recital 71 and the guidelines also task companies with preventing discrimination in many forms,

[88] Several U.S. scholars have called for algorithmic due process that closely mirrors what is in the GDPR. *See, e.g.*, Citron, *supra* note 1; Citron & Pasquale, *supra* note 1; Crawford & Schultz, *supra* note 1.

[89] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 27 (should assess "all the relevant data").

[90] *Id.* at 32.

[91] *Id.*

[92] Kaminski, *Binary Governance*, *supra* note 3, at 34.

[93] Malgieri & Comandé, *supra* note 8, at 258–59. GDPR, *supra* note 6, Recital 71 states that companies should adopt "technical and organisational measures appropriate to ensure . . . that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised."

[94] *See* GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 32.

including on the basis of race, ethnic origin, political opinion, religion.[95] The guidelines envision ongoing testing and feedback into an algorithmic decision-making system to prevent errors, inaccuracies, and discrimination on the basis of sensitive ("special category") data.[96]

As for whether Member States are bound to create laws incorporating these same safeguards—that is, whether the GDPR harmonizes safeguards against algorithmic decision-making or leaves space for Member State variations—the guidelines are strongly suggestive but not entirely clear. They state that "Member . . . State law that authorizes [algorithmic decision-making] must also incorporate appropriate safeguarding measures."[97] In the next paragraph, the guidelines state that "[s]uch measures should include as a minimum a way for the data subject to obtain human intervention, express their point of view, and contest the decision."[98] This suggests that the GDPR does harmonize safeguards, even when a Member State creates a new exception to the ban on automated decision-making. But as several scholars point out, Member State laws have already developed variations on Article 22's safeguards.[99]

To return to the larger claim: while the guidelines and Recitals do not eliminate all room for interpretation, they largely clarify the GDPR's algorithmic accountability provisions to make them more, not less, rigorous. These interpretive documents fully close a number of the loopholes suggested by scholars and limit room for others. This causes Article 22 (and accompanying notification and access rights) to be broader, stronger, and deeper than the preceding EU algorithmic accountability regime.[100] The GDPR applies to more activity (is broader), comes with more significant

[95] *See id.* at 6, 10, 14 (explaining that even in profiling without automated decision-making, companies should employ "safeguards aimed at ensuring fairness, non-discrimination and accuracy in the profiling process"); *see also* GDPR, *supra* note 6, Recital 71 ("[P]revent, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or processing that results in measures having such an effect.").

[96] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 28.

[97] *Id.* at 27.

[98] *Id.* at 27.

[99] *See* Brkan, *supra* note 8, at 12 (describing German law on insurance); *see also* Malgieri, *supra* note 41, at 8-9 (describing variations in Member State laws as to suitable safeguards for algorithmic decision-making).

[100] Amy Kapczynski has used similar terms ("broader," "deeper," and "more severe") to describe the ratcheting-up of intellectual property law internationally. Amy Kapczynski, *The Access to Knowledge Mobilization and the New Politics of Intellectual Property*, 117 YALE L.J. 804, 821 (2008).

enforcement (is stronger), and adds significant protections (is deeper), compared to the Data Protection Directive.

Article 22 applies to or restricts more activity, and is, therefore, broader than Article 15 of the Data Protection Directive. Where the DPD's provisions were limited to automated decision-making connected to individual profiling—that is, processing for the purpose of "evaluat[ing] certain personal aspects" of the person—Article 22 is not limited to profiling.[101] Automated decision-making may often "partially overlap with or result from profiling[,]"[102] but the guidelines make clear that Article 22's scope goes beyond personal profiling to other kinds of automated decisions.[103]

Article 22 is also broader by virtue of being interpreted to apply to decisions involving human rubber-stamping, where several Member States had interpreted the Directive's provisions to apply only to automated decisions involving no human at all.[104] Similarly, where some Member States implemented the DPD's provisions as a right to object, the Working Party guidelines explain that Article 22 is a prohibition on algorithmic decision-making.[105] It thus applies to all automated decision-making, not just when an individual voices an objection. Thus several of the interpretations advanced by the Working Party ensure that Article 22 will apply to more activity than the DPD did.

Second, Article 22 is stronger than the Directive's provisions, meaning that it is harder law.[106] The GDPR provides both stronger penalties and stronger enforcement mechanisms.[107] And where Member States could change the

[101] Mendoza & Bygrave, *supra* note 3, at 10, 11; GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 8 ("Automated decision-making has a different scope and may partially overlap with or result from profiling . . . Automated decisions can be made with or without profiling; profiling can take place without making automated decisions.").

[102] *See* GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 8.

[103] *See id.* (discussing example of imposing speeding fines based on evidence from speed cameras).

[104] *See* GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 21; Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist*, *supra* note 8, at 94–95.

[105] *See* GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 19.

[106] *See, e.g.*, Kenneth W. Abbott et al., *The Concept of Legalization*, 54 INT'L ORG. 401, 404 (2000) (describing a spectrum of "legalization" along the three dimensions of obligation, precision, and delegation); Kal Raustiala & Anne-Marie Slaughter, International Law*, International Relations and Compliance*, *in* THE HANDBOOK OF INTERNATIONAL RELATIONS 538, 552 (Thomas Risse & Beth Simmons eds., 2002).

[107] *See, e.g.*, Casey et al., *supra* note 8, at 165–70.

wording and in practice the meaning of the DPD through implementation, the GDPR, as a regulation, has direct effect within Member States. Thus, the wiggle room in Article 22 is lessened (even as the text still contemplates some variations by Member States) and the enforcement authority behind it is greatly strengthened.

Finally, Article 22's protections run deeper than the DPD's provisions. Specifically, the mandatory requirements for companies are more significant under the GDPR than they were under the DPD. Under the DPD, if the contract exception applied, it was not clear that a company needed to do anything else to protect individual rights—it need not necessarily adopt safeguards.[108] By contrast, Article 22 requires safeguards—even when an exception applies—that, at a minimum, include a right to human intervention, a right to object, and a right to express one's view.[109] As discussed above, the Working Party guidelines and Recitals clarify that these measures include both an individual right to explanation and multiple systemic accountability requirements such as audits.

Article 22 and the accompanying notification and access provisions in Articles 13, 14, and 15 thus put in place an algorithmic accountability regime that is broader, stronger, and deeper than the largely symbolic regime that existed under the DPD. Accompanied by other company duties in the GDPR—including establishing data protection officers, using data protection impact assessments, and following the principles of data protection by design—this regime, if enforced, has the potential to be a sea change in how algorithmic decision-making is regulated in the EU.[110]

[108] *See* Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, 1995 O.J. (L 281) art. 15(2)(a)

> [I]s taken in the course of the entering into or performance of a contract, provided the request for the entering into or the performance of the contract, lodged by the data subject, has been satisfied or that there are suitable measures to safeguard his legitimate interests, such as arrangements allowing him to put his point of view.

[109] *See* GDPR, *supra* note 6, at art. 22(3).

[110] *See* Casey et al., *supra* note 8, at 173–88 (describing data protection impact assessments and data protection by design and by default); *see also* Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 23, 68–80 (identifying the GDPR's actual algorithmic accountability regime as consisting of DPIAs, PbD, and other individual GDPR rights such as the right to erasure).

V. THE RIGHT TO EXPLANATION, REVISITED

Against this backdrop of the GDPR's strengthened algorithmic accountability regime, this Article now returns to the much-debated right to explanation. Transparency is a basic principle of the GDPR.[111] In fact, it can be striking to a U.S. audience just how many of the GDPR's rights resemble open government laws, rather than traditional privacy causes of action.[112] This is because data protection regimes are grounded in fairness, and transparency and fairness are linked ideals; we often use transparency as an element of accountability, to establish that systems are fair.[113] But in the right to explanation debate, the centrality of transparency to the GDPR has gotten lost. Several scholars have, pessimistically, vastly underrepresented what kinds of disclosures about algorithmic decision-making are required under the GDPR.[114] To be fair, these scholars largely wrote before the Working Party guidelines were finalized. But now that the final version of the guidelines has been released, some explanation of explanation is overdue.

To understand what is at stake, it is worth briefly summarizing the back-and-forth over transparency that has taken place in the literature. Scholars on both sides of the Atlantic have called for transparency in algorithmic decision-making, in the form of both notice towards individuals and audits that enable expert third-party oversight.[115] Some of these calls for transparency have been

[111] *See* GDPR, *supra* note 6, at art. 5(1)(a); GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 9 ("Transparency of processing is a fundamental requirement of the GDPR.").

[112] *Compare, e.g.*, GDPR, *supra* note 6, at arts. 12–15, *with* the U.S. Privacy Act, 5 U.S.C. § 552(a) (comparing the GDPR's rights of transparency, notification, and access to the U.S. Privacy Act, which provides individual rights of transparency into public systems of records). *Compare, e.g.*, the GDPRs implementation of the Fair Information Practice Principles (FIPS), *with* the Prosser privacy torts. For an overview of the FIPS, see GDPR, *supra* note 6, at art. 5. *See also* GDPR, *supra* note 6, Recital 39. For a discussion of the Prosser torts, see Neil M. Richards & Daniel J. Solove, *Prosser's Privacy Law: A Mixed Legacy*, 98 CALIF. L. REV. 1887, 1891–903 (2010).

[113] Robert Gellman, *Fair Information Practices: A Basic History* (Apr. 10, 2017) (unpublished manuscript) (explaining the principles of transparency and fairness that are at the base of worldwide data protection regimes); *see* ORG. FOR ECON. CO-OPERATION & DEV., THE OECD PRIVACY FRAMEWORK 15 (Sept. 23, 1980, *revised* 2013) ("Openness Principle" and "Individual Participation Principle": "An individual should have the right: a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him").

[114] *See, e.g.*, Edwards & Veale, *Slave to the Algorithm*, *supra* note 8; Wachter et al., *Counterfactual*, *supra* note 8; Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist*, *supra* note 8.

[115] *See, e.g.*, Citron, *supra* note 1, at 1305; Citron & Pasquale, *supra* note 1; Crawford & Schultz, *supra* note 1; Hildebrandt, *supra* note 1; Kim, *supra* note 1; FRANK PASQUALE, THE

ambitiously deep and broad, suggesting that both algorithmic source code and data sets should be subjected to public scrutiny.[116] Others have responded by enumerating the harms this level of transparency could cause,[117] or by arguing that transparency directed at individuals will be relatively useless since individuals lack the expertise to do much with it.[118] But transparency of some kind has a clear place in algorithmic accountability governance, from recent calls for algorithmic impact assessments to proposals for whistleblower protections, to regularly repeated calls for algorithmic auditing.[119]

The GDPR comes closest to creating what Frank Pasquale has called "qualified transparency": a system of targeted revelations of different degrees of depth and scope aimed at different recipients.[120] Transparency in practice is not limited to revelations to the public.[121] It includes putting in place internal company oversight, oversight by regulators, oversight by third parties, and

BLACK BOX SOCIETY 140–88 (2015) (calling this "qualified transparency"—"limiting revelations in order to respect all the interests involved in a given piece of information").

[116] Citron *supra* note 1, at 1308; Citron & Pasquale, *supra* note 1, at 20, 26 (the "logics of predictive scoring systems should be open to public inspection"). Citron & Pasquale also note that information about the datasets (but not the datasets themselves) could be released to the public. *Id.* at 27 (noting that Zarsky says the public could be informed about datasets without social risk); Zarsky, *supra* note 1, at 1563.

[117] Ananny & Crawford, *supra* note 1, at 978 ("[F]ull transparency can do great harm."); Kroll et al., *supra* note 1, at 639; Zarsky, *supra* note 1, at 1553–63.

[118] Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 64, 67 ("Individuals are mostly too time-poor, resource-poor, and lacking in the necessary expertise to meaningfully make use of these individual rights."); Kroll et al., *supra* note 1, at 638 ("The source code of computer systems is illegible to nonexperts.").

[119] *See, e.g.*, Desai & Kroll, *supra* note 1 (calling for whistleblower protections); A. Michael Froomkin, *Regulating Mass Surveillance as Privacy Pollution: Learning from Environmental Impact Statements*, 2015 U. ILL. L. REV. 1713, 1713 (2015) (calling for "requirements for those conducting mass surveillance in and through public spaces to disclose their plans publicly via an updated form of environmental impact statement"); Price, *supra* note 1, at 421 (arguing that the FDA should pursue a "more adaptive regulatory approach with requirements that developers disclose information underlying their algorithms"); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109 (2017) (describing the potential benefits of "algorithmic impact statements [requiring] police departments to evaluate the efficacy and potential discriminatory effects of all available choices for predictive policing technologies"); David Wright & Charles D. Raab, *Constructing a Surveillance Impact Assessment*, 28 COMPUTER L. & SECURITY REV. 613 (2012) (describing "surveillance impact assessment (SIA), a methodology for identifying, assessing and resolving risks . . . posed by the development of surveillance systems").

[120] PASQUALE, *supra* note 115, at 142.

[121] *See* Zarksy, *supra* note 1, at 1532 ("Intuitively, transparency is linked to merely one meaning—that the relevant information is disseminated broadly to (1) the *general public*" but "[f]ully understanding this concept, however, calls for distinguishing among the *recipients* of the information transparency policy provides."). *But see* Kroll et al., *supra* note 1, which appears to define transparency only as disclosure to the public.

communications to affected individuals. Each of these revelations may be of a different depth or kind; an oversight board might get access to the source code, while an individual instead might get clearly communicated summaries that she can understand.

To summarize the right to explanation and accompanying transparency measures, as some have, as a "transparency fallacy"—palliative measures requiring mere icons or simplistic explanations—is to both misrepresent their actual substance and mischaracterize the GDPR's overall transparency regime.[122] The GDPR's individual transparency provisions are deeper than some have suggested. And the overall accountability regime that the GDPR puts in place establishes multiple layers of transparency, some of which go very deep indeed. This Part starts with individual transparency rights, before turning to the systemic approach to algorithmic accountability that the GDPR puts in place.

Individuals have a "right to be informed" about algorithmic decision-making.[123] That right is housed both in the "meaningful information about the logic involved" provisions of Articles 13 and 14 and in Article 22(3)'s suitable safeguards provision.[124] It is true that the guidelines state that individuals need not be provided with source code or complex mathematical explanations, under either Article 22 or the accompanying notification and access provisions.[125] But that is because those individual transparency provisions are meant to serve the purpose of providing expert oversight.

The "who" and "why" of transparency in the GDPR dictates the what, when, and how. Individual transparency provisions, as the guidelines make clear, are intended to empower individuals to invoke their other rights under the GDPR.[126] Therefore, while individuals need not be provided with source code, they should be given far more than a one-sentence overview of how an algorithmic decision-making system works. They need to be given enough information to be able to understand what they are agreeing to (if a company

[122] Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 43; Wachter et. al, *Counterfactual*, *supra* note 8, at 865–66, 887.

[123] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 20.

[124] *Id.* at 20, 25 ("Providing this information will also help controllers ensure they are meeting some of the required safeguards referred to in Article 22(3) and Recital 71.").

[125] *Id.* at 25 ("[N]ot necessarily a complex explanation of the algorithms used or disclosure of the full algorithm."), 31 ("Instead of providing a complex mathematical explanation about how algorithms or machine-learning work, the controller should consider using clear and comprehensive ways to deliver the information to the data subject.").

[126] *Id.* at 27 ("The controller should provide the data subject with general information . . . which is also useful for him or her to challenge the decision . . . . The data subject will only be able to challenge a decision or express their view if they fully understand how it has been made and on what basis.").

is relying on the explicit consent exception);[127] to contest a decision;[128] and to find and correct erroneous information, including inferences.[129]

Scholars have (in this Article's view, disingenuously) suggested that the GDPR's transparency requirements in Article 12—requirements that companies make an effort to communicate information in a way understandable to individuals— restrict the depth and quality of information a company must reveal.[130] Article 12 demands that companies communicate clearly, to ensure that individuals can in fact act on the information they receive. It aims to prevent companies from flooding individuals with useless or unnecessarily complicated or time-wasting information, abusing notice requirements to create obscurity through information floods. [131] In other words, Article 12 requires that companies make their communications to individuals comprehensible. It does *not* reduce the GDPR's substantial disclosure requirements to meaninglessly high-level or simplistic information

[127] *Id.* at 13 ("Controllers seeking to rely upon consent as a basis for profiling will need to show that data subjects understand exactly what they are consenting to.").

[128] *Id.* at 27.

[129] *Id.* at 17–18 ("Individuals may wish to challenge the accuracy of the data used and any grouping or category that has been applied to them. This rights to rectification and erasure apply to both the 'input personal data' (the personal data used to create a profile), and the 'output data' (the profile itself or 'score' assigned to the person)."), 31 ("Controllers providing data subjects with access to their profile in connection with their Article 15 rights should allow them the opportunity to update or amend any inaccuracies in the data or profile.").

[130] *See* Wachter et. al, *Counterfactual*, *supra* note 8, at 865 ("Detailed information appears to not be necessary as Art. 12(7) states that the required information can be provided along with standardi[z]ed icons . . . proposed icons reveal the initial expectations of regulators for simple, easily understood information."), 866 ("[E]ach provision suggests that information disclosures need to be tailored to their audience, with envisioned audiences including children and uneducated laypeople."), 887 (illustrating simplistic transparency infographics that were ultimately not adopted by the European Parliament, and stating that these "reveal the level of complexity expected by EU legislators" in an explanation to a data subject and that "[t]he reliance on generic icons suggests that individual-level, contextualised information is not required").

[131] *See* Ananny & Crawford, *supra* note 1, at 979 ("[*S*]*trategic opacity*—in which actors 'bound by transparency regulations' purposefully make so much information 'visible that unimportant pieces of information will take so much time and effort to sift through that receivers will be distracted from the central information the actor wishes to conceal.' "); Zarsky, *supra* note 1, at 1508 ("The process of merely flooding the public with facts and figures does not effectively promote transparency. It might even backfire."); *see also* Wendy E. Wagner, *Administrative Law, Filter Failure, and Information Capture*, 59 DUKE L.J. 1321, 1324–25 (2010); GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 31 ("Instead of providing a complex mathematical explanation . . . the controller should consider using clear and comprehensive ways to deliver the information to the data subject.").

or infographics. Companies can be required to communicate in-depth information at the same time that they are required to communicate it clearly.[132]

Communication to individuals about algorithmic decision-making must thus be simultaneously understandable (or "legible"),[133] meaningful, and actionable. It must be understandable to individuals, rather than delivered in complex jargon or as an information flood.[134] However, it must also convey considerable depth; the guidelines note that "[c]omplexity is no excuse for failing to provide information."[135] And it must provide enough information that an individual can act on it—to contest a decision, or to correct inaccuracies, or to request erasure.[136]

Thus, there is a clear relationship between the other individual rights the GDPR establishes—contestation, correction, and erasure—and the kind of individualized transparency it requires. This suggests something interesting about transparency: the substance of other underlying legal rights often determines transparency's substance.[137] If one has a right of correction, one needs to see errors. If one has a right against discrimination, one needs to see what factors are used in a decision. Otherwise, information asymmetries render underlying rights effectively void.

The guidelines list examples of what kinds of information should be provided to individuals and how it should be provided. Individuals should be told both the categories of data used in an algorithmic decision-making process and an explanation of why these categories are considered relevant.[138]

[132] *See, e.g.*, RANDALL MUNROE, THING EXPLAINER: COMPLICATED STUFF IN SIMPLE WORDS (2015) (Munroe "used line drawings and only the thousand (or, rather, "ten hundred") most common words to provide simple explanations for some of the most interesting stuff there is"). Thanks to Matthew R. Cushing for the pointer.

[133] Malgieri & Comandé, *supra* note 1, at 245 (introducing the concept of legibility to this debate: "legibility is concerned with making data and analytics algorithms both transparent and comprehensible") (citing Richard Mortier, et al., *Human Data Interaction: The Human Face of the Data-Driven Society*, MIT TECH. REV. (2014); *see* Zarsky, *supra* note 1, at 1520 (discussing the related concept of interpretability).

[134] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 31 ("[C]lear and comprehensive").

[135] *Id.* at 25, n.40.

[136] *See id.* at 17, 27, 31; *see also* Mendoza & Bygrave, *supra* note 3, at 16 (explaining that the possibility of a "right of ex post explanation of automated decision is implicit in the right 'to consent' a decision"); Selbst & Powles, *supra* note 8, at 242 (explaining that enhancing data subject rights to include the right to "contest a decision" is reinforced by "GDPR's emphasis on meaningful transparency . . . in a way that is useful, intelligible, and actionable to the data subject").

[137] Selbst & Barocas, *Intuitive Appeal*, *supra* note 3, at 1120–21.

[138] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 31 (explaining good practice recommendations for data controllers).

Moreover, they should be told the "factors taken into account for the decision-making process, and . . . their respective 'weight' on an aggregate level . . . ."[139] They should be told how a profile used in algorithmic decision-making is built, "including any statistics used in the analysis[,]"[140] and the sources of the data in the profile.[141] Lastly, companies should provide individuals an explanation of why a profile is relevant to the decision-making process and how it is used for a decision.[142]

The GDPR's individualized system of algorithmic transparency thus requires far more than a counterfactual explanation (e.g., "if you were not 25, you would have gotten this job").[143] The guidelines further note, in several places, that companies should use technological design to create more effective notice mechanisms, such as through "visuali[z]ation and interactive techniques."[144] Not only is it a company's duty to communicate a particular depth of information, but a company must also pay attention to using effective design choices to ensure that information is both noticed and understood.

This does not mean that the individual right to explanation and the accompanying transparency rights in the GDPR give individuals a right to all information about an algorithm. Nor does it mean to suggest that the conversation about what information must be released to individuals ends here. It is clear from the guidelines that this conversation will be ongoing. There is still room to read in, for example, a best practice of releasing performance metrics, which the guidelines do not suggest.[145] Two scholars have proposed a number of suggestions of the kind of information that would be useful—including both information about the model (the family of model, training parameters, summary input data, human-understandable averages of how inputs become outputs, how the model was tested, trained, or screened) and information about the individual decision (counterfactuals, which cases

[139] *Id.* at 27 ("[W]hich is also useful for him or her to challenge the decision.").

[140] *Id.* at 31.

[141] *Id.*

[142] *See id.*

[143] *But see* Wachter et. al, *Counterfactual*, *supra* note 8.

[144] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 31 ("Controllers may want to consider implementing a mechanism for data subjects to check their profile, including details of the information and sources used to develop it."). *Id.* at 32 ("Controllers could consider introducing online preference management tools such as a privacy dashboard."). Hildebrandt, *supra* note 1, at 53 (calling for "TETs": transparency-enhancing tools); Citron & Pasquale, *supra* note 1, at 29 (suggesting interactive modeling).

[145] Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 55; Malgieri & Comandé, *supra* note 8, at 259.

are most similar to the individual's, what characteristics cause individuals to receive similar treatment, how confident the system is of a specific outcome).[146]

But the GDPR's individual algorithmic transparency rights, accompanied by other GDPR transparency rights, go a long way towards establishing what U.S. scholars have called for—including revealing the sources of data, inferences about an individual, and even some math.[147] Throughout, the emphasis is on individual understanding of information of a meaningful depth, so that an individual subject of algorithmic decision-making can invoke her rights.

Other forms of systemic transparency that go substantially deeper accompany this individualized transparency regime. Individuals might not have access to source code or datasets, but other parties do. The GDPR's regime of systemic transparency is established through Article 22's safeguards provision and the Working Party interpretation of it, and through more general GDPR provisions such as the requirement of impact assessments.[148] This systemic transparency regime includes the requirement of data protection impact assessments for automated processing, the general information-forcing and oversight powers granted to regulatory authorities.

There are a number of ways that systematic transparency can be implemented. First, regulators can use significant information-forcing capabilities under the GDPR to get access to information about algorithms.[149] The GDPR also envisions general data protection audits conducted by government authorities.[150]

Second, most companies deploying algorithmic decision-making must set up internal accountability and disclosure regimes. They must perform a data protection impact assessment,151 and provide information to an internal but

[146] Edwards & Veale, *Slave to the Algorithm*, *supra* note 8, at 55–56, 58.

[147] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 31 (mentioning "any statistics used in the analysis").

[148] *Id.* at 28, 32 (discussing safeguards under art. 22). GDPR, *supra* note 6, at art. 35, art. 58.

[149] GDPR, *supra* note 6, at art. 58(1)(e) (authorizing the authority to carry out data protection audits, and "obtain, from the controller and the processor, access to all personal data and to all information necessary for the performance of its tasks").

[150] *See id.* at art. 58(1)(b).

[151] *Id.* at art. 35(3)(a) (requiring a data protection impact assessment "in a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person"); GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 29–30 (explaining that this requirement "will apply in the case of decision-making including profiling

independent data protection officer who has, at least on paper, deep information-forcing abilities.[152] Companies that fall under Article 22 must also give human reviewers deeper transparency onto "all the relevant data" as part of the right to human intervention.[153]

Third, the guidelines suggest that companies performing decision-making with a "high impact on individuals" should use independent third-party auditing and provide that auditor with "all necessary information about how the algorithm or machine learning system works."[154] Hence, the GDPR's approach to systemic accountability establishes a second aspect of Pasquale's "qualified transparency": deeper information flows, including source code, both within companies and to regulatory authorities and third-parties. It is true that this information does not get released to the public. But it is myopic to focus only on the individual version of transparency and decry its shallowness, rather than seeing its place and purpose in a system of required information flows.

The purpose of each transparency measure affects not just the depth of information revealed but also the timing of transparency.[155] Discrete events in the GDPR trigger individual transparency—when, for example, data is collected,[156] a decision is made,[157] an individual's consent is obtained,[158] or an individual requests information.[159] This connects individualized transparency to the rights of an individual, but limits the efficacy of individualized transparency at creating oversight over the construction of an algorithm, or its ongoing performance. In particular, individual transparency rights largely occur after the fact of algorithmic development, when it is far more difficult (if not impossible) to impose accountability or corrections on a system.[160] By

with legal or similarly significant effects that is not wholly automated, as well as solely automated decision-making defined in Article 22(1)").

[152] *Id.* at art. 38(2) ("The controller and processor shall support the data protection officer in performing the tasks . . . by providing . . . access to personal data and processing operations . . . ."); GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 29–30.

[153] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 27 (assess "all the relevant data").

[154] *Id.* at 32.

[155] Ananny & Crawford, *supra* note 1, at 982 (discussing the "temporal dimension of transparency").

[156] *See* GDPR, *supra* note 6, at arts. 13, 14.

[157] *See id.* at art. 22; GDPR, *supra* note 6, Recital 71.

[158] GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 12–13.

[159] GDPR, *supra* note 6, at art. 15

[160] Kroll et al., *supra* note 1 at 659–60, 662; Desai & Kroll, *supra* note 1 at 39–42.

contrast, the GDPR's systemic accountability measures are envisioned as ongoing, continuous,[161] and being implemented early on in an algorithm's development. This creates, in theory at least, internal, expert/third-party, and regulatory oversight over the development of an algorithm from its inception, better serving the purposes of correcting error, inaccuracy, and bias in a changing system over time.

VI. CONCLUSION

The GDPR sets up a system of "qualified transparency" over algorithmic decision-making that gives individuals one kind of information, and experts and regulators another. This multi-pronged approach to transparency should not be dismissed as lightly as some have done. There is an individual right to explanation. It is deeper than counterfactuals or a shallow and broad systemic overview, and it is coupled with other transparency measures that go towards providing both third-party and regulatory oversight over algorithmic decision-making. These transparency provisions are just one way in which the GDPR's system of algorithmic accountability is potentially broader, deeper, and stronger than the previous EU regime.

It is one thing to put these requirements on paper and quite another to have them operate in practice. The system of algorithmic accountability that the GDPR and its accompanying interpretative documents envision faces significant hurdles in implementation: high costs to both companies and regulators, limited individual access to justice, and limited technical capacity of both individuals and regulators. As I note elsewhere, there are other ways in which the GDPR may fail.[162] Its heavy reliance on collaborative governance in the absence of significant public or third-party oversight could lead to capture or underrepresentation of individual rights.[163]

But for companies with a footprint in the EU, it is important to note that the GDPR does govern algorithmic decision-making, and many of the

[161] *See, e.g.*, GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING, *supra* note 9, at 28

> Controllers should introduce appropriate procedures and measures to prevent errors, inaccuracies or discrimination on the basis of special category data. These measures should be used on a cyclical basis; not only at the design stage, but also continuously, as the profiling is applied to individuals. The outcome of such testing should feed back into the system design.

See, e.g., Ananny & Crawford, *supra* note 1, at 976.

[162] *See* Kaminski, *Binary Governance*, *supra* note 3, at 67–68.

[163] *See, e.g.*, CHRISTINA ANGELOPOULOS ET AL., STUDY OF FUNDAMENTAL RIGHTS LIMITATIONS FOR ONLINE ENFORCEMENT THROUGH SELF-REGULATION (2016) (discussing the problems raised by delegating individual rights protection to companies).

potential loopholes in that system have been limited or closed. Companies face a decision of whether to put humans meaningfully back in the loop of algorithmic decision-making and thus escape Article 22. Otherwise, they must put in place a significant set of safeguards, including both individual rights and ongoing internal and third-party accountability measures.

ECONOMIC ANALYSIS OF NETWORK EFFECTS AND INTELLECTUAL PROPERTY

Peter S. Menell[†]

ABSTRACT

The information revolution has brought demand-side effects to the fore of economic activity, business strategy, and intellectual property jurisprudence and policy. Intellectual property doctrines play a central role in harnessing network effects, promoting innovation to overcome excess inertia, and balancing consumer welfare, competition, and innovation. This Article surveys and integrates the economic, business strategy, and legal literatures relating to network effects and intellectual property. Part I introduces the topic of network effects and provides an overview of the Article. Part II describes the functioning of network markets. Part III examines the interplay of business strategy, contract, standard setting organizations, intellectual property, and competition policy. Part IV presents three principles for tailoring intellectual property regimes and competition policy to network technologies. Part V traces the evolution of intellectual property protection for network features of systems and platforms. Part VI discusses the interplay of intellectual property protection and competition policy. Part VII assesses the extent to which intellectual property protection and competition policy align with the normative design principles. Part VIII identifies promising areas for future research.

DOI: https://doi.org/10.15779/Z381V5BD7T

[†] Koret Professor of Law and Director, Berkeley Center for Law & Technology, University of California at Berkeley School of Law. Thanks to participants at the 2015 Economics of Intellectual Property Research Handbook Conference, the 2017 Intellectual Property Scholars Conference, and especially Michael Carrier for their comments on this project. I am grateful to Alex Barata, Concord Cheung, Louise Decoppet, Amit Elazari, Andrea Hall, Megan McKnelly, Reid Whitaker, and Samantha Vega for research assistance.

TABLE OF CONTENTS

I. INTRODUCTION

The economics of intellectual property begins with the classic appropriability problem: In a competitive economy, imitators can enter markets for information goods after inventors and authors have incurred research and development (R&D) costs and sell the innovative or creative product at the cost of reproduction. Without means for appropriating an adequate return on investment in R&D, the market will under-produce technological advances and creative expression.[1]

The provision of intellectual property protection for technological advances and creative expression affords inventors and authors a mechanism to recoup their investments, although not without imposing the deadweight loss of monopoly exploitation and potentially interfering with cumulative creativity.[2] Conventional analysis of intellectual property seeks to optimize the duration and scope of intellectual property rights in order to balance these tradeoffs.[3] This framework applies to goods and services for which consumer demand is independent—i.e., where one consumer's utility from consuming a good or service does not depend on choices of other consumers.

Yet consumer demand for information goods and services can be interdependent, especially in the digital age. The consumers' valuation of systems technologies—such as telecommunication networks (e.g., telephone networks, cable systems, satellite systems, and Internet protocols), interconnected devices (e.g., mobile phones, operating systems and application programs, printers and replacement cartridges, and audio-video devices and media), databases (e.g., Internet searches), and electric charging stations (e.g., Tesla superchargers)—often depends upon other consumers' choices. For example, a smartphone platform with many adopters will attract more app developers, thereby increasing the functionality and value of that platform for consumers, developers of complementary goods (e.g., apps), and the platform sponsor.

[1] *See* Peter S. Menell & Suzanne Scotchmer, *Intellectual Property Law*, *in* 2 HANDBOOK OF LAW AND ECONOMICS 1473, 1478–79, 1499–500 (A. Mitchell Polinsky & Steven Shavell eds., 2007).

[2] *See* PETER S. MENELL, MARK A. LEMLEY & ROBERT P. MERGES, INTELLECTUAL PROPERTY IN THE NEW TECHNOLOGICAL AGE: 2018, VOL I: PERSPECTIVES, TRADE SECRETS, AND PATENTS 168 (2018).

[3] *See* Nancy T. Gallini, *Patent Policy and Costly Imitation*, 23 RAND J. ECON. 52, 62–63 (1992) (analyzing "optimal patent design when costly imitation displaces a patentee's output as the length of patent protection increases"); *see generally* WILLIAM D. NORDHAUS, INVENTION, GROWTH, AND WELFARE: A THEORETICAL TREATMENT OF TECHNOLOGICAL CHANGE (1969).

Platforms function like a common language. Devices that "speak" a common language (such as a programming language, application program interface (API) specification, or a set of graphical user icons) can communicate with other devices and humans familiar with that language. Innovators can more easily design peripheral equipment that expands the functionality of existing devices. Over time, users internalize how a computer language or application program represents functions, often memorizing the most commonly used series of keystrokes or developing macros customized to perform their most common tasks. These human capital investments commit users to particular languages and platforms and encourage employers to adopt systems that are widely known by prospective employees so as to recruit promising candidates and reduce training costs.[4] Thus, it is common for people seeking jobs in programming, accounting, and design fields to list those computer languages and application programs that they have mastered on their resumes. Network externalities arise from the enhanced labor mobility and reduced training costs produced by shared, or at least compatible, computer systems across different work environments. When people in different places can communicate more efficiently through compatible file formats, network externalities result.

The value of networks grows disproportionately with their adoption bases. Such positive feedback dynamics drive a growing number of markets in the information economy,[5] from computer operating systems to mobile phones, printers (and ink cartridges), video game consoles, Internet search engines (such as Google), Internet commerce (such as eBay and Amazon), social networks (such as Facebook, LinkedIn, and Tinder), cloud computing, the Internet of Things, and shared economy platforms (such as Airbnb and Uber).

Advances in digital and network technologies have dramatically reshaped the competitive and innovative landscape. As a consultant for the Internet dating industry has remarked, "[i]t's never been cheaper to start a dating site and never been more expensive to grow one."[6] Dating apps usually start by offering free services to new users, seeking to build a viral bandwagon. If they gain traction through innovative features or marketing, they then face the daunting task of monetizing the network, typically through advertising or

[4] *See* Neil Gandal, *Hedonic Price Indexes for Spreadsheets and an Empirical Test for Network Externalities*, 25 RAND J. ECON. 160, 168–69 (1994).

[5] *See generally* CARL SHAPIRO & HAL R. VARIAN, INFORMATION RULES: A STRATEGIC GUIDE TO THE NETWORK ECONOMY 103–226 (1999).

[6] Alina Tugend, *For Online Dating Sites, a Bumpy Road to Love*, N.Y. TIMES (Dec. 24, 2016), https://www.nytimes.com/2016/12/24/business/online-dating-sites-jdate-christianmingle.html [https://perma.cc/3AME-4DKS].

membership fees. Monetization, however, can reverse the positive feedback effects, thereby reducing the network's size, unraveling the network's benefits, and jeopardizing the platform's sustainability. Finding the right balance between viral growth and monetization is the principal challenge of a growing range of enterprises in the Internet Age.

The interdependence of consumer demand has important ramifications for the design of intellectual property and competition policy. In a static economic model (i.e., one without innovation), consumers benefit from robust competition within product standards. Open access to product standards encourages realization of network externalities. Although bandwagon effects can enhance consumer welfare in a static context, they can also make it more difficult for developers of improved platforms to enter the market. Consumers and suppliers of complementary products can face significant switching costs in migrating from one platform to another. For example, once businesses have invested heavily in developing programs to run on a software platform (e.g., macros for the Lotus 1-2-3 spreadsheet), it becomes much more difficult for a competitor offering an enhanced spreadsheet (e.g., Borland Quattro Pro) to enter the market unless they can provide a low-cost migration path. Facebook's widespread success and user investment made it difficult for even Google to build a sustainable competing social network. Orkut, Google Buzz, Google Friend Connect, and Google+ have failed or languished.

The technical standards governing access to platforms, commonly referred to as application program interfaces (APIs) in the software industry, play a critical role in consumer and programmer adoption decisions, market entry, and competition. Those who control a widely-adopted platform can obstruct new innovative platforms and complementary products and services (such as refilling and repair). Familiarity with the user interface and features, connections to other network adopters (such as Facebook friends), and investments in complementary assets (such as macros that run on the platform) can keep consumers on an otherwise inferior platform. The human capital investment in learning an API can lock programmers into a platform, and sunk costs in manufacturing facilities, fabrication designs, and contracts with suppliers and customers can lock manufacturers into design choices.

At the same time, the ability to secure an innovative platform can be vital to investing in the R&D needed to advance systems technologies. Without the prospect of earning a significant return on research, development, and marketing of a new platform, investors have little incentive to take on the risk of investing the substantial resources necessary to challenge an entrenched platform. Therefore, the availability, scope, and remedies for intellectual property protection for network features of systems technologies and platforms (e.g., interface specifications) provide a key strategic asset for controlling network markets and a critical mechanism for promoting advances in network technologies.

Demand-side or network effects, therefore, complicate the design of an optimal intellectual property regime. Control of interface specifications and other network features of computer technologies through intellectual property protection has become the key to market dominance in a growing number of important Information Age markets. Nearly all of the major software copyright disputes, as well as a key exception to the Digital Millennium Copyright Act's (DMCA) anti-circumvention provisions, have revolved around the protectability of interface specifications. Patent protection for network technologies has also become a critical battleground with some disputes centered on licensing network technologies through standard setting organizations. Trade secrecy, trademark protection, and contract law are also important tools for regulating competition in network markets.

This Article explores the critical role of intellectual property in network markets as well as the ramifications of network effects for the design of intellectual property regimes. Part II describes the functioning of network markets. Part III examines the interplay of business strategy, contract, standard setting organizations, intellectual property, and competition policy with regard to network coordination. Part IV presents three principles for tailoring intellectual property regimes and competition policy for network technologies. Part V traces the evolution of intellectual property protection for network features of systems and platforms. Part VI discusses the interplay of intellectual property protection and competition policy. Part VII assesses the extent to which intellectual property protection and competition policy align with normative design principles. Part VIII identifies promising areas for future research.

II. FUNCTIONING OF NETWORK MARKETS

In many market settings, consumers' utility functions are independent. Take, for example, the market for ice cream. My enjoyment of a particular flavor (e.g., hazelnut chocolate chip), style (e.g., gelato), or brand (e.g., Talenti) does not depend significantly on the utility that other consumers derive from the purchase and consumption of ice cream. It is possible that greater popularity of a flavor, style, or brand makes that combination more widely available or lowers the price due to economies of scale on the production side, but competition usually ensures efficient allocation of resources in these circumstances. The effects are more likely pecuniary, which work through the market and only affect the distribution of value, than technological, which affects the economic efficiency of the economy.[7]

By contrast, some market equilibria depend critically on the number of consumers that have joined or are likely to join a particular platform. Take, for example, a social network like Facebook. A new entrant to this market, say Google+, might offer enhanced functionality. But if most of my social network is already on Facebook and I cannot easily bridge the two networks, then I am far less likely to switch.

Network effects have long been central to human civilization and market economies. Languages, measurement systems (metric versus imperial), electrical equipment standards (alternating current versus direct current; computer networking protocols), driving conventions (left side versus right side), and railroad gauges (the width between and across rails) are notable examples where demand-side coordination greatly influences consumer welfare, economic efficiency, and social discourse. Standardized railroad gauge, for example, supported far-reaching railroad networks, promoted competition in locomotive and railcar markets, and enabled interconnected rail services.[8] Part III focuses on how such coordination or standardization occurs through business strategy, technological innovation, intellectual property law, industry and consumer coordination, and government policies (including antitrust law).

The economic and social value of network effects can be substantial. According to Metcalfe's Law—attributed to Robert Metcalfe, co-inventor of the Ethernet, a local computer network platform that foreshadowed and ushered in the Internet—the value of a telecommunications network is proportional to the square (n^2) of the number of devices (or nodes (n)) in the system. This economic "law" reflects the potential number of contacts within a network and assumes that they are each of equal value. Even though this theoretical maximum is unlikely to be obtained in the real world,[9] the powerful growth potential of network systems drives much of the information economy. The net value, of course, also depends on the cost per user. In many telecommunications and computer applications, such costs are low and have declined over time because of Moore's Law—Intel co-founder Gordon Moore's audacious, yet remarkably accurate, prediction that the number of

[7] *See* Stan J. Liebowitz & Stephen E. Margolis, *Network Externality: An Uncommon Tragedy*, 8 J. ECON. PERSP. 133, 137–39 (1994).

[8] *See* Douglas J. Puffert, *The Standardization of Track Gauge on North American Railways, 1830-1890*, 60 J. ECON. HIST. 933, 944–47 (2000).

[9] *See* Bob Briscoe, Andrew Odlyzko & Benjamin Tilly, *Metcalfe's Law Is Wrong*, IEEE SPECTRUM (July 1, 2006), https://spectrum.ieee.org/computing/networks/metcalfes-law-is-wrong [https://perma.cc/ET4W-D7YV].

transistors on an integrated circuit would double every two years (later reduced to eighteen months).[10]

Both real and virtual networks can produce these effects.[11] Real networks entail physical connectivity enabling a user to interact or communicate directly with others. They include transportation systems (such as railroad gauges), telecommunication systems (such as a telephone or broadcast network), and media systems (such as data storage devices). By contrast, virtual networks operate through the evolution of markets for complementary products. The supply of complementary goods typically drives these markets. For example, by enabling programmers to develop apps for the iOS platform, Apple promotes a virtual network surrounding its iPhone and other computer devices. The availability of apps on iOS drives demand for iOS devices, which in turn attracts app developers. More apps generate a wide range of functionality, thereby spurring increased demand for iPhones. Other examples of virtual networks include application programs that enable users to share data files with other programs and users, ATM cards and automatic teller machines, credit cards, and the merchants who accept them, and next generation payment systems such as Apple Pay and Square. The defining feature of virtual networks is that the demand for the product depends significantly on the availability of complementary goods and services.

The magnitude of network effects depends on several considerations: interdependencies of consumer utility functions, range of complementary products or services, availability of alternative platforms, switching costs, business strategies, and legal limits on leveraging network markets (such as intellectual property protection and competition policy). In some cases, physical limitations govern network access—e.g., where a device must physically or digitally interoperate with other devices. In others, the network is not physically constrained, but instead driven by consumer familiarity or ease of use.

The design determinants of a network market—interoperability or compatibility standards—are shaped by the type and degree of ownership, sponsorship, and governance of network access. Some network standards are

[10] *See* Gordon E. Moore, *Cramming More Components onto Integrated Circuits*, 38 ELECTRONICS 114 (1965) (predicting that the number of transistors in an integrated circuit would double approximately every two years); *see also* Jonathan Borwein & David H. Bailey, *Moore's Law Is 50 Years Old but Will It Continue?*, CONVERSATION (July 20, 2015), http://theconversation.com/moores-law-is-50-years-old-but-will-it-continue-44511 [https://perma.cc/8PFQ-Q84J].

[11] *See* SHAPIRO & VARIAN, *supra* note 5, at 183; *see also* Michael Katz & Carl Shapiro, *Network Externalities, Competition, and Compatibility*, 75 AM. ECON. REV. 424, 424–25 (1985).

established or authorized by a government, international organization, or formal standard-setting organization (SSO). These are sometimes referred to as de jure standards as they have official backing and can be enforced by law. Such enforcement can limit or afford access to standards. Individual companies or consortiums sponsor many important network standards. These are sometimes referred to as de facto standards, although they might be backed by patent, copyright, trademark, or false advertising law.

An important distinction in network markets relates to whether a standard is "free," open, closed (i.e., proprietary), or somewhere in the middle.[12] The Free Software Movement allows other users to run, study, share, copy, and modify the software so long as these users permit use of any derivative works on the same terms. "Open source" software typically connotes that the software or interface is freely available to any market participant, but there might or might not be restrictions on the availability of complementary goods embodying the standard. A closed or proprietary standard is one in which a sponsoring enterprise or organization regulates access, typically through licensing of intellectual property rights.

The distinction between open and closed standards can be ambiguous. For example, many SSOs require that participating enterprises license standard-essential patents on fair, reasonable, and non-discriminatory (FRAND) terms.[13] Substantial uncertainty arises because patent owners rarely fully specified in advance which patents are "standard-essential" or the license terms on which they are available.[14] On the other hand, "free" software licensed pursuant to the General Public License (GPL) requires users to make available any software incorporating the licensed code under the same "share and share alike" restriction.[15]

The controversy over the Java API platform illustrates the complexity that can arise surrounding intermediate—i.e., partially open—platforms.[16] Sun Microsystems released the Java programming language without restriction in

[12] *See* HEATHER J. MEEKER, OPEN (SOURCE) FOR BUSINESS: A PRACTICAL GUIDE TO OPEN SOURCE SOFTWARE LICENSING 31–47 (2015) (describing the economic and technological forces that shape standards); *see also* Joel West, *The Economic Realities of Open Standards: Black, White, and Many Shades of Gray*, *in* STANDARDS AND PUBLIC POLICY 87 (Shane Greenstein & Victor Stango eds., 2007).

[13] *See generally* Michael Mattioli, *Patent Pool Outsiders*, 33 BERKELEY TECH. L.J. 233 (2018).

[14] *See* Jorge L. Contreras & Richard J. Gilbert, *A Unified Framework for RAND and Other Reasonable Royalties*, 30 BERKELEY TECH. L.J. 1451 (2015).

[15] *See* Brian W. Carver, *Share and Share Alike: Understanding and Enforcing Open Source and Free Software Licenses*, 20 BERKELEY TECH. L.J. 443, 455 (2005).

[16] *See* Peter S. Menell, *Rise of the API Copyright Dead: An Updated Epitaph for Copyright Protection of Network and Functional Features of Computer Software*, 31 HARV. J.L. & TECH. 305, 346–414 (2018); *see also* Mark A. Lemley & David McGowan, *Could Java Change Everything? The Competitive Propriety of a Proprietary Standard*, 43 ANTITRUST BULL. 715, 756–72 (1998).

part to prevent Microsoft from leveraging its Windows desktop computer operating system monopoly into dominance of website functionality. Sun's Java strategy promoted the "Write Once, Run Anywhere" (WORA) principle: the notion that any browser can execute Java applets (small application programs, such as those used for animated web pages) on any operating system—including on Microsoft Windows, Unix, macOS, and Linux.

Over time, Sun developed pre-written API packages to facilitate Java programming. Sun developed the Java Community Process (JCP), a quasi-public formalized administrative process, for developing technical specifications for Java technology and extensions. Sun used the JCP and licensing of the Java trademark to promote collaboration and commitment to the WORA principle. When Google sought to use some, but not all, of the Java APIs to develop the Android platform and licensed Android using a less restrictive licensing regime (i.e., not requiring that derivative works be shared on a "free" basis), Sun and Oracle (which acquired Sun Microsystems in 2010) objected, resulting in one of the costliest intellectual property battles in recent memory.[17]

Network effects arise whenever the value that consumers place on a product or service depends upon the number of other consumers or programmers purchasing that product or using that service. As the number of adopters (or the installed base) of a platform grows, the benefits of being part of that platform increase. For example, consumers generally prefer telephone networks or protocols offering the largest user bases.

Like economies of scale (declining unit costs with increased production) on the supply-side of a market, the value of a network generally increases with widespread adoption. The availability of better application programs to run on an operating system platform will lead more consumers to prefer that operating system, which in turn will spur a greater quantity and quality of application programs for that operating system. Whereas economies of scale typically fall off at some point due to technical or organizational limits, positive feedback on the demand side generally continues to increase with the size of the installed base. For this reason, a single standard or a very small number of standards are likely to predominate in markets with strong network effects, as reflected in Microsoft's dominance in the market for microcomputer operating systems, Google's dominance among Internet search engines, and Facebook's dominance as a social network.

[17] *See* Menell, *supra* note 16. Section V.B.5 explores this litigation.

The high value that consumers place upon standardization, however, can make it particularly difficult for improved products to break into the market. Such bandwagon effects can stifle development and diffusion of improved technology platforms.[18]

III. INTERPLAY OF BUSINESS STRATEGY, CONTRACT, STANDARD SETTING, INTELLECTUAL PROPERTY, AND COMPETITION POLICY

The dynamics of network technologies produce a particularly complex strategic playing field. Firms typically choose among three strategies when competing in network markets: (1) market dominance through establishing and controlling a new proprietary standard; (2) adopting an existing standard either through imitation (where it is legally permissible) or licensing; or (3) working with other firms in the industry—either informally, contractually, through formal industry organizations, or through governmental standardization bodies—to develop an open or quasi-open standard.[19]

Among the strategies firms use to establish their product or service as the de facto industry standard are: massive advertising campaigns; penetration pricing (pricing products or services below cost or giving them away in order to hasten adoption by consumers); issuing impressive product preannouncements to entice consumers and discourage competitors; providing adopters with various forms of insurance (such as short-term leases or pricing arrangements that tie the price of the system to the number of adopters); licensing of the product in order to grow the network more rapidly (and to create competition in the expansion of the network); and vertical integration and strategic investments into markets for complementary products to assure consumers that valuable application programs will be available.[20]

[18] *See* Joseph Farrell & Garth Saloner, *Standardization, Compatibility, and Innovation*, 16 RAND J. ECON. 70, 75–79 (1985).

[19] *See* Joseph Farrell & Garth Saloner, *Coordination Through Committees and Markets*, 19 RAND J. ECON. 235 (1988); Stanley M. Besen & Joseph Farrell, *Choosing How to Compete: Strategies and Tactics in Standardization*, 8 J. ECON. PERSP. 117 (1994); *see generally* SHAPIRO & VARIAN, *supra* note 5, at 184–86; Joseph Farrell & Timothy Simcoe, *Choosing the Rules for Consensus Standardization*, 43 RAND J. ECON. 235 (2012).

[20] *See generally* SHAPIRO & VARIAN, *supra* note 5; Farrell & Simcoe, *supra* note 19; Joseph Farrell & Paul Klemperer, *Coordination and Lock-In: Competition with Switching Costs and Network Effects*, *in* 3 HANDBOOK OF INDUSTRIAL ORGANIZATION (Mark Armstrong & Robert H. Porter eds., 2007); Kenneth C. Baseman, Frederick R. Warren-Boulton & Glenn A. Woroch, *Microsoft Plays Hardball: The Use of Exclusionary Pricing and Technical Incompatibility to Maintain Monopoly Power in Markets for Operating System Software*, 40 ANTITRUST BULL. 265, 273–80 (1995)

Adopting an existing standard enlarges the size of a network comprising both the entrant's product and its rival's—the existing platform's—products. This increases the desirability of the rival's products to consumers, thereby reducing the adopter's market share (although of a larger market) relative to what it would have been had the firm adopted an incompatible product standard. Thus, even though the net social welfare of adopting a rival's standard may exceed the net social welfare of introducing an incompatible standard, the entrant may nonetheless prefer to adopt an incompatible standard because the entrant cannot appropriate all the benefits of compatibility, some of which accrue to past and present purchasers of the rival's products.[21]

Firms often pursue Strategy 2 (adopting an existing standard) and Strategy 3 (collaborating with other firms in establishing a standard) in tandem. Both strategies create a more traditional market setting in which firms compete over price, quality, and services to win market share on a common platform. This achieves greater competition on a particular platform and fosters the realization of network externalities but may impair competition to innovate better platforms.[22] The market dominance strategy is often riskier but can produce the highest payoff for the winner.

A firm's strategy will depend on a range of factors, including its reputation among consumers for serving the type of network market that it has targeted, its available resources (and access to capital markets) to make the investments in distribution and marketing necessary to persuade consumers that the firm will prevail in the standard battle, the strength of its technology for establishing a standard (although such technology need not be superior to others on the market), and complementary assets within the firm or strong strategic alliances in vertical markets. The firm's strategy will also depend upon the availability of intellectual property protection, contractual means, and technological controls (e.g., encryption technology) for precluding, limiting, or delaying access by competitors to the firm's standard.

(describing Microsoft's monopolistic pricing strategies); Joseph Farrell & Garth Saloner, *Installed Base and Compatibility: Innovation, Product Preannouncements, and Predation*, 76 AM. ECON. REV. 940, 940 (1986).

[21] *See* Katz & Shapiro, *supra* note 11, at 435 (finding that firms with good reputations or large existing networks might pursue a proprietary strategy even when social welfare is increased by purusing a compatibility statetgy because the firms cannot appropriate the full value (or enough of the value) of the network externalities); *see generally* Michael L. Katz & Carl Shapiro, *Technology Adoption in the Presence of Network Externalities*, 94 J. POL. ECON. 822 (1986).

[22] *See* Michael L. Katz & Carl Shapiro, *Systems Competition and Network Effects*, 8 J. ECON. PERSP. 93, 108–10 (1994).

IBM successfully pursued the market dominance strategy when it entered the microcomputer market in the early 1980s. IBM combined its reputation for serving the mainframe market and technological and marketing capabilities with copyright and trade secrecy protection for its basic instruction operating system (BIOS) chip. IBM's strategic hold on the industry quickly unraveled, however, when competitors successfully reverse engineered the BIOS chip,[23] making much less expensive, fully compatible IBM clones available on the market by the mid to late 1980s. IBM exited the microcomputer hardware industry soon thereafter.

Microsoft emerged as the winner in the microcomputer industry during this upheaval. Its DOS operating system, on which IBM had previously built its microcomputers, emerged as the de facto standard. Robust competition in microcomputers using DOS and a growing array of application programs (including several Microsoft flagship products such as Word and later Excel) drove adoption of DOS-based computers and fueled Microsoft's dominance. Microsoft skillfully migrated users from DOS to Windows, withstanding Apple's assertion of intellectual property control of the Mac desktop graphical user interface. By the mid-1990s, Microsoft dominated the microcomputer industry through its control of the Windows platform. Apple was a distant second and fading.

The emergence of the Internet in the mid-1990s opened new modes of competition in computer markets. Netscape's Navigator Internet browser and Sun's highly interoperable Java platform threatened Microsoft's dominance in the microcomputer and software marketplace.[24] Microsoft responded by integrating its browser technology, Internet Explorer, into the Windows operating system and engaging in restrictive licensing agreements with microcomputer manufacturers, thereby reducing the effective price of its browser to zero. Consequently, the market for Netscape's browser evaporated. Microsoft also undermined Java's efforts to establish a universal meta-platform for software application programs by offering a proprietary, non-interoperable version.[25]

Network effects have allowed one or a few firms to dominate many Internet markets, including search (Google), social networks (Facebook),

[23] *See* Russell Moy, *A Case Against Software Patents*, 17 SANTA CLARA COMPUTER & HIGH TECH. L.J. 67, 70–73 (2000) (chronicling reverse engineering of the IBM BIOS); *see generally* Peter S. Menell, *Tailoring Legal Protection for Computer Software*, 39 STAN. L. REV. 1329 (1987) (analyzing legal protection for computer software).

[24] *See* Lemley & McGowan, *supra* note 16, at 741–42.

[25] *See* Sun Microsystems, Inc. v. Microsoft Corp., 999 F. Supp. 1301, 1305 (N.D. Cal. 1998).

mobile (iOS, Android), commerce (Amazon, eBay), content streaming (YouTube, Netflix, Spotify), payment systems (PayPal), and sharing networks (Airbnb, Uber). Apple successfully regained prominence in critical digital markets through its mobile and App Store network market strategies.

Formal standardization plays a tremendous role in many electronics and telecommunications markets.[26] Russell traces electrical standardization through formal standard-setting organizations for more than a century.[27] These processes have relied on engineers and scientists seeking to promote the best engineering solutions to technical challenges. They form key infrastructure for the electronics and telecommunications industries. Engineers from major technology companies participate in dozens of standard-setting organizations, including many of the leading professional engineering societies, such as the IEEE Standards Association and the Internet Engineering Task Force (IETF). These processes have carried over to semiconductor designs, mobile phones, Internet protocols, and computer devices. A typical laptop computer today embodies more than 250 technical standards.[28]

Intellectual property protection for network technologies can significantly influence the development of standards, follow-on innovation, and market competition. Patents in the information and communication technology fields (semiconductors, computers, and mobile phones) have presented the most salient concerns.

Building on Williamson's classic treatment of economic holdup[29]—whereby asymmetric information, transaction costs, and incomplete contracts create the potential for a contracting party to extract the value of sunk or locked in, relationship-specific investments—Lemley and Shapiro[30] posit a patent bargaining model in the shadow of strong potential remedies (automatic injunctive relief and large monetary awards) that generates an analogous

[26] *See* Jorge L. Contreras, *Technical Standards, Standards-Setting Organizations and Intellectual Property: A Survey of the Literature (with an Emphasis on Empirical Approaches)*, *in* 2 RESEARCH HANDBOOKS ON THE ECONOMICS OF INTELLECTUAL PROPERTY LAW (Peter S. Menell & David L. Schwartz eds., 2019); Mark A. Lemley, *Intellectual Property Rights and Standard-Setting Organizations*, 90 CALIF. L. REV. 1889, 1894–95 (2002).

[27] *See* Andrew L. Russell, *Industrial Legislatures: The American System of Standardization*, *in* INTERNATIONAL STANDARDIZATION AS A STRATEGIC TOOL 71, 72–76 (2006).

[28] *See* Contreras, *supra* note 26, at 7.

[29] *See generally* OLIVER WILLIAMSON, THE ECONOMIC INSTITUTIONS OF CAPITALISM (1985).

[30] *See* Mark A. Lemley & Carl Shapiro, *Patent Holdup and Royalty Stacking*, 85 TEX. L. REV. 1991 (2007).

inefficient dynamic. Companies that unwittingly sink large investments into infringing products are subject to having such investments extracted through patent infringement litigation. Such extraction can greatly exceed the contribution of the patented technology relative to the best non-infringing alternative. The presence of multiple patents covering a single product—what has been referred to as the patent thicket problem[31]—exacerbates holdup effects, creating a royalty stacking problem: total patent royalty demands may exceed the contribution of patented technologies to the market demand for the product.

Various scholars have questioned Lemley and Shapiro's assumptions and empirical basis for royalty stacking.[32] They note that royalty stacking is unlikely to occur with full information and low transaction costs. There is good reason, however, to question optimism about ex-ante bargaining. Ziedonis, for example, finds that firms acquire patents more aggressively when the patents for numerous component technologies of an industry—like the semiconductor industry—are widely distributed.[33] The proliferation of patent litigation over information and communication technology indicates that intellectual property protection imposes at least some implicit tax on these network industries. Nonetheless, more recent empirical research raises doubts about the severity of royalty stacking. Galetovic and Gupta, for example, find that mobile wireless prices have fallen, quantities have grown, and the industry has become less concentrated over time, indicating that royalty stacking may not be as serious as prior research had claimed.[34] Barnett surveys the growing

[31] *See* Carl Shapiro, *Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting*, 1 INNOVATION POL'Y & ECON. 119, 119–22 (2000).

[32] *See* Einer Elhauge, *Do Patent Holdup and Royalty Stacking Lead to Systematically Excessive Royalties?*, 4 J. COMPETITION L. & ECON. 535 (2008); J. Gregory Sidak, *Holdup, Royalty Stacking, and the Presumption of Injunctive Relief for Patent Infringement: A Reply to Lemley and Shapiro*, 92 MINN. L. REV. 714 (2008); Damien Geradin, Anne Layne-Farrar & A. Jorge Padilla, *The Complements Problem Within Standard Setting: Assessing the Evidence on Royalty Stacking*, 14 B.U. J. SCI. & TECH. L. 144 (2008); Damien Geradin & Miguel Rato, *Can Standard-Setting Lead to Exploitative Abuse? A Dissonant View on Patent Hold-Up, Royalty Stacking and the Meaning of FRAND*, 3 EUROPEAN COMPETITION J. 101 (2015); John M. Golden, *"Patent Trolls" and Patent Remedies*, 85 TEX. L. REV. 2111 (2007). *But see* Mark A. Lemley & Carl Shapiro, *Reply: Patent Holdup and Royalty Stacking*, 85 TEX. L. REV. 2163 (2007).

[33] *See* Rosemarie H. Ziedonis, *Don't Fence Me In: Fragmented Markets for Technology and the Patent Acquisition Strategies of Firms*, 50 MGMT. SCI. 804, 817–19 (2004).

[34] *See* Alexander Galetovic & Kirti Gupta, *Royalty Stacking and Standard Essential Patents: Theory and Evidence from the World Mobile Wireless Industry* 24–25 (Hoover Institution Working Grp. on Intellectual Prop., Innovation & Prosperity, IP Working Paper Series No. 15012, Mar. 2017).

literature and concludes that the evidence of royalty stacking is weak.[35] All would agree, however, that industry coordination through patent pooling and SSOs can alleviate these problems.[36] Such pools, however, can facilitate collusion, raise barriers to entry, and spark other public policy concerns.[37]

Notwithstanding the widespread use of standard-setting processes and agreements on technical standards, the rules governing access to standards and the licensing of patented technologies are rarely specified in advance. Standard setting organizations (SSOs) exercise caution to avoid violating antitrust laws barring price-fixing. In addition, many companies participating in standard-setting processes do not wish to reveal their patent prosecution strategies or pre-commit to price terms. Thus, most technical SSOs require only that participants disclose their patented technologies and agree to license standard-essential patents (SEPs) on FRAND terms. The potential for holdup and royalty stacking remains.[38]

Some sectors of the software industry have alleviated or avoided these risks by committing to open source policies.[39] Viral forms of open source licensing, such as the GPL, however, can discourage investment in downstream innovation by limiting direct appropriability for technological advances. For this reason, Google chose a more permissive open source license for Android.[40] This fostered collaboration and rapidly expanded the Android network while encouraging innovation by handset makers and telecommunications companies.

[35] *See* Jonathan M. Barnett, *Has the Academy Led Patent Law Astray?*, 32 BERKELEY TECH. L.J. 1313,1344–61 (2017).

[36] *See generally* NAT'L RESEARCH COUNCIL OF THE NAT'L ACADS., PATENT CHALLENGES FOR STANDARD-SETTING IN THE GLOBAL ECONOMY: LESSONS FROM INFORMATION AND COMMUNICATIONS TECHNOLOGY (Keith Maskus & Stephen A. Merrill eds., 2013).

[37] *See generally* Richard J. Gilbert, *Antitrust for Patent Pools: A Century of Policy Evolution*, 2004 STAN. TECH. L. REV. 3 (2004).

[38] *See* Contreras, *supra* note 26, at 16–20; *see generally* FED. TRADE COMM'N, THE EVOLVING IP MARKETPLACE: ALIGNING PATENT NOTICE AND REMEDIES WITH COMPETITION (2011) [hereinafter FTC REPORT]; Lemley & Shapiro, *supra* note 30.

[39] *See* Robert P. Merges, *A New Dynamism in the Public Domain*, 71 U. CHI. L. REV. 183 (2004).

[40] *See* Menell, *supra* note 16, at 357–72.

IV. RAMIFICATIONS FOR INTELLECTUAL PROPERTY AND COMPETITION POLICY

As the preceding analysis suggests, intellectual property protection can play a critical role in network markets. As one software entrepreneur metaphorically explained, creating an API is analogous to building a city:

> First you try to persuade applications programmers to come and build their businesses on [your tract of land]. This attracts users, who want to live there because of all the wonderful services and shops the programmers have built. This in turn causes more programmers to want to rent space for their businesses, to be near the customers. When this process gathers momentum, it's impossible to stop.
>
> Once your city is established, owning the API is like being the king of the city. The king gets to make the rules: collecting tolls for entering the city, setting the taxes that the programmers and users have to pay, and taking first dibs on any prime locations (by keeping some APIs confidential for personal use).[41]

This Part discusses the general economic considerations bearing on whether and to what extent intellectual property ought to protect network features of systems technologies—those features that affect access to or interoperability with a system. It also presents three principles for tailoring intellectual property regimes and competition policy for network technologies.

There are two market failures in play in optimizing intellectual property protection. First, network features of system technologies, like any other technology, are subject to the classic appropriability problem. Without intellectual property protection, inventors of more advanced platform technologies will be subject to being undercut by new entrants who imitate the innovations without bearing R&D costs. First-mover advantages, effective marketing, trade secrecy, and other strategies might provide sufficient motivation for some R&D, but there is reason to be concerned that the unregulated market will under-produce potentially high value, but risky and costly, innovation in network technologies.

Demand-side effects in network markets, however, complicate the conventional analysis of intellectual property protection. Because of the dynamics of network markets, some firms might be motivated to limit access to their platforms to reap the outsize profits from controlling a network market. This strategy, however, can hinder the realization of network benefits by raising prices, limiting access by third parties, and discouraging innovation

[41] *See* JERRY KAPLAN, STARTUP: A SILICON VALLEY ADVENTURE 49–50 (1995) (explaining that "our value is the APIs" and "the real wars [in the computer industry] are over control of APIs" (quoting an industry remark)).

because of the high barriers to entry. Consumers benefit when they and their devices, systems, and programs "speak" the most widely adopted platform—the lingua franca—or can translate that code into language their devices understand. This often provides for greater functionality, such as more software that will run on their platform and larger communication networks.

Second, widely adopted product standards can strand the industry on an obsolete platform.[42] Consumers resist switching costs—from learning new tools and languages to acquiring new devices. They demand substantial improvements in efficiency or functionality to jettison comfortable, well-worn devices and software tools for new tools and systems.

Thus, the installed base built upon the dominant platform—reflected in durable goods and human capital (training) specific to the old standard—can create inertia that makes it much more difficult for any one producer to break away from the prevailing standard by introducing a noncompatible product, even if the new standard offers a significant technological improvement over the current standard.[43] In this way, network externalities can retard innovation and slow or prevent adoption of improved product standards.

Therefore, companies seeking to leapfrog a widely adopted standard face substantial risk. They must not only invent an improved platform, but they must also devise and execute a successful strategy to migrate consumers from the dominant platform. They also face the challenge of encouraging other software and complementary product developers to build for the new platform. One strategy is to steeply discount the costs of the new platform or provide free access. This strategy is not sustainable unless the platform developer has ancillary revenue streams—such as bundled advertising or ties to other products and services—to cover their research, development, product, and support costs.

Intellectual property protection can contribute to *and* alleviate the network externality dilemma. On the one hand, intellectual property protection for the network features of computer technology can discourage realization of positive network externalities by limiting access to network technologies. The sponsor of a particular network technology can use intellectual property protection to exclude competitors or charge a high licensing fee for access, thereby raising costs. The intellectual property owner can also limit innovation by restricting how the network technology evolves. On the other hand, intellectual property protection can provide valuable incentives for

[42] *See generally* Farrell & Saloner, *supra* note 18.

[43] *See generally* Farrell & Saloner, *supra* note 20.

overcoming bandwagon effects that entrench obsolete standards.[44] Without the potential for a large reward, inventors contemplating innovative new platforms might not be willing to make the substantial, risky R&D and marketing investments needed to challenge, and hopefully leapfrog, the incumbent platform.

These considerations suggest three principles for intellectual property protection of APIs and other functional features of platform technologies: (A) a parsimony principle to prevent firms from establishing protection for product standards without providing a significant technological advance; (B) a proportionality principle to ensure that firms can appropriate a fair return on technological advances in platform innovation sufficient to overcome the excess inertia of network markets, but not so large as to stunt network externalities; and (C) a deterrence principle to discourage deceptive practices and overreach in network markets.

A. PARSIMONY PRINCIPLE: NO INTELLECTUAL PROPERTY PROTECTION FOR FUNCTIONAL ATTRIBUTES ABSENT SIGNIFICANT TECHNOLOGICAL ADVANCE

Consumers benefit from access to platforms that produce network benefits. Those benefits can increase over time through positive feedback effects and the development of aftermarket enhancements and complementary products. The incentives for firms adopting product standards, however, are distorted. New entrants might choose an incompatible standard to differentiate their products from established brands, even where growing the established network would enhance consumer welfare.[45]

Intellectual property protection affects such choices by setting the ground rules for establishing proprietary platforms. Firms will be more inclined to build competing platforms where the thresholds for acquisition of intellectual property protection—and hence the power to exclude subsequent entrants and those seeking to bridge platforms—are low.

Thus, intellectual property regimes should discourage platform adoption choices that undermine realization of network externalities unless there is a large countervailing benefit, such as substantial technological advance. Affording meaningful intellectual property protection for network technologies without requiring a significant technological advance encourages wasteful differentiation and increases the risk of undeserved monopoly power. With easy access to intellectual property protections—for example, by merely using arbitrary lock-out codes—firms can fragment platforms that would

[44] *See* Menell, *supra* note 23, at 1343.

[45] *See* Katz & Shapiro, *supra* note 11, at 425, 434–36; Katz & Shapiro, *supra* note 21, at 822, 830–33.

otherwise foster competition in the non-network product features and in downstream products competing on the platform. Through serendipity, first mover advantage, clever marketing, or simply luck, market power can emerge through positive feedback effects without discernible consumer benefits. Therefore, intellectual property law should not simply reward novel (but obvious) or expressive functional features of network goods or services. Rather, strong protection should be reserved for substantial advances.

B. PROPORTIONALITY PRINCIPLE: OVERCOMING EXCESS INERTIA WITHOUT UNDUE PROTECTION

While low thresholds for intellectual property protection for network technologies undermine realization of network externalities, balanced protection for substantial technological advances may be necessary for entrants to overcome the strong inertial forces driving network markets. Switching costs discourage consumers from making the leap to a new platform. For network products and services, those costs can be particularly high due to network effects. The leap is likely not worth the cost for modest technological improvements. At some point, however, overall consumer welfare will be enhanced by migration to an alternative platform. The efficient tipping point depends on R&D and marketing costs as well as the contours of consumer demand.

The excess inertia of network effects can hinder, delay, and possibly prevent the technological shift to a substantially more advanced technological platform. If all such advances were freely available to entrants, the free-rider problem would discourage the R&D and marketing investment needed to displace the obsolete platform. Yet providing strong intellectual property protection for such advances can lead to robust returns as the market tips to the new platform.

The shift from "feature phones"—mobile phones "featuring" voice and text messaging with rudimentary Internet access—to true "smartphones" with email and robust web functionality illustrates the challenges and opportunities surrounding network markets. Through the 1990s, Motorola, Nokia, and a few other vendors established the first generation of mobile devices. Sun, Microsoft, and Symbian vied to establish the platform for mobile devices that integrated email and Internet capabilities. By 2005, Java's Micro Edition (ME) was faring well, with adoption by Palm and Blackberry.

As the first-generation smartphone battle was resolving, Apple was secretly investing heavily in an ambitious new platform. Intellectual property played a significant role in motivating Apple's R&D. As Steve Jobs noted during the historic January 2007 iPhone announcement, "boy have we patented it!"[46]

Meanwhile, Google was at work on its own skunkworks[47] smartphone play: the Android smartphone platform. Given Google's concern that its success in search and online advertising could be displaced if Microsoft or Apple gained dominance in the shift to mobile devices, Google sought to develop an open platform that would perpetuate Google's dominance in search and other services on mobile devices.[48]

This standards war illustrates the dynamism of network markets as well as the complex role of intellectual property protection. In the space of just a few years, the market shifted dramatically from feature phones to rudimentary smartphones and then to advanced smartphones. By 2011, Apple and Google dominated the market. Intellectual property protection played a central role in encouraging investment, but also resulted in massive resources devoted to intellectual property acquisition, coalition building, standard setting on upstream technologies, and litigation.

There is no simple answer to the question of how much protection is enough, especially given the range of business strategies, institutions, and intellectual property regimes that can deliver appropriate returns on investment, the dynamism of network markets, and concerns about anti-competitive leveraging network technology dominance. Lichtman emphasizes strong property rights to promote platform competition,[49] but this analysis assumes low transaction costs, overlooks consumers' cognitive limitations stemming from lock-in, and risks leveraging monopoly power and inhibiting cumulative innovation.

The optimal level of intellectual property protection has a dynamic quality, with the level of protection dissipating as network technologies and platforms become dominant. Menell recommends a limited patent-type regime to protect the functional features of computer software, although with shorter duration

[46] Tim Worstall, *Too Funny, Steve Jobs Invalidated an Apple Patent over Prior Art*, FORBES (Sept. 27, 2013), https://www.forbes.com/sites/timworstall/2013/09/27/too-funny-steve-jobs-invalidated-an-apple-patent-over-prior-art/ [https://perma.cc/4HDR-ZU3U].

[47] *See* FRED VOGELSTEIN, DOGFIGHT: HOW APPLE AND GOOGLE WENT TO WAR AND STARTED A REVOLUTION 45 (2013). "Skunkworks" was derived from Lockheed's code-named secret World War II project to develop a new fighter jet ("Skunk Works"), which was taken from Al Capp's Li'l Abner comic strip, a "skunkworks" project brings together a small group of highly skilled researchers to pursue radical innovations. *See* Menell, *supra* note 16, at 347–48.

[48] *See* Menell, *supra* note 16, at 357.

[49] *See* Douglas Lichtman, *Property Rights in Emerging Platform Technologies*, 29 J. LEGAL STUD. 615, 615–20 (2000).

and more flexibility to promote access to platforms that become widely adopted.[50] Menell[51] advocates a genericide-type doctrine,[52] which could protect emerging platforms but give way to broader access when a platform becomes dominant and risks affording the proprietor the ability to leverage that control to hinder cumulative innovators.[53] This analysis anticipated Microsoft's rise and its abusive market tactics in undermining Netscape and Sun. At the same time, scholars have opposed copyright protection for the functional and interoperable aspects of computer technology so as to avoid large returns to first movers that win a standards battle without offering significant technological innovation.[54] Such limitations on copyright protection afford competitors freedom to use and build on unpatented methods of operation. In some circumstances, compulsory licensing of patents might be desirable. This can be achieved through injunctive relief.

The proportionality principle ensures that platform innovators who choose proprietary strategies (as opposed to more collaborative approaches) have the potential to reap significant rewards if they prevail in a standards competition, but that their ability to control the platform (and charge monopoly prices) declines as the network becomes entrenched. Such a regime creates optimal conditions for overcoming excess inertia while promoting the realization of network benefits. It also allows for competition to enhance and improve established platforms.

C. Deterrence Principle: Discouraging Overreach with Balanced Remedies

Intellectual property law and competition policy should also protect against deceptive practices and leveraging intellectual property rights to control network markets. The integrity of standard-setting processes is particularly

[50] *See* Menell, *supra* note 23.

[51] *See* Peter S. Menell, *An Analysis of The Scope of Copyright Protection for Application Programs*, 41 STAN. L. REV. 1045, 1101–04 (1989).

[52] A trademark can become generic and thereby lose protection if it becomes associated in the public's mind with a category of product rather than the source of a particular brand of the product. *See, e.g.*, Murphy Door Bed Co., Inc. v. Interior Sleep Sys., Inc., 874 F.2d 95 (2d Cir. 1989) ("Murphy bed" for a bed that folds up into a wall cabinet); King-Seeley Thermos Co. v. Aladdin Indus., 321 F.2d 577 (2d Cir. 1963) ("Thermos" for a vacuum insulated bottle).

[53] *See* Richard N. Langlois, *Technological Standards, Innovation, and Essential Facilities: Toward a Schumpeterian Post-Chicago Approach*, *in* DYNAMIC COMPETITION AND PUBLIC POLICY: TECHNOLOGY, INNOVATION, AND ANTITRUST ISSUES 193 (Jerry Ellig ed., 2001).

[54] *See* Menell, *supra* note 23; Dennis S. Karjala, *Copyright, Computer Software, and the New Protectionism*, 28 JURIMETRICS 33, 62–72 (1987); Pamela Samuelson, Randall Davis, Mitchell D. Kapor & J. H. Reichman, *A Manifesto Concerning the Legal Protection of Computer Programs*, 94 COLUM. L. REV. 2308, 2332–64 (1994).

critical to efficient collaboration among enterprises and innovators working in network industries. The choice of standards depends on a range of factors, including potential restrictions on practicing technological standards. Hence, standard-setting bodies should require disclosure of all potential intellectual property encumbrances or, at a minimum, advance commitment by SSO members to licensing such technologies on fair and reasonable terms. Courts should penalize efforts to reduce transparency in standard-setting processes and take failure to abide by such commitments into consideration in enforcing patent rights.

Antitrust law and competition policy should also take network effects into account in assessing monopoly power, scrutinizing collaborations and contractual agreements, and fashioning remedies. The consumer, competitive, and innovation ramifications of network markets are especially complex. What might appear to be benign and welfare-improving behaviors—such as integrating a "free" browser into an operating system product or bundled after-market services—might ultimately lead to monopolization of important emerging and downstream markets. Hence, antitrust law must be vigilant in assessing the dynamism and path-dependence of network technologies. For example, advance determination of licenses for standard-essential patents can promote competition in downstream products and services. In some circumstances, antitrust authorities should tolerate some collusive behaviors—such as ex-ante negotiation of FRAND license rates by SSOs—that resemble forbidden price-setting. The Sherman Antitrust Act bars contracts and conspiracies that *unreasonably* restrain competition. In network markets, some collaboration promotes economic efficiency.

The crafting of remedies to combat abusive and anti-competitive behavior in network markets requires careful consideration of effects on consumers and competitors. Once a standard has taken root and is generating substantial network benefits, traditional remedies—such as enjoining the offensive activities or breaking up dominant firms—can cause adverse effects on the consumers who have adopted the standard as well as other downstream users—such as programmers and competitors who have incurred sunk costs in joining the platform. Leveraging intellectual property rights to control network markets might also produce countervailing innovative efficiencies. Hence, antitrust authorities and courts should consider remedies that promote the realization of network benefits while also promoting enhanced competition and innovation. In some circumstances, these considerations favor compulsory licenses, which can be flexible and adaptable, over injunctive remedies.

V. INTELLECTUAL PROPERTY PROTECTION FOR NETWORK FEATURES

In view of the tremendous economic significance of controlling access to systems technologies by exploiting demand-side effects and excluding competition in complementary goods and services, such as repair services, replacement parts, and ancillary markets (e.g., advertising and consumer data), platform developers and entrepreneurs have sought to use intellectual property to protect APIs and other means to exclude competitors from their platforms and systems. As an alternative approach, computer programmers and a growing number of commercial enterprises in the open-source community have deployed intellectual property protection as a tool for sharing technology and precluding proprietary control of core Internet and computer operating system technologies.

Since the principal forms of intellectual property protections developed long before the advent of digital technology, which made network effects so important, the intellectual property statutes do not expressly reflect the aforementioned policy principles for APIs and other functional features of platform technologies. Nonetheless, the mixed statutory/common law heritage of intellectual property law[55] has afforded courts discretion to interpret statutory provisions, adapt common law doctrines, and apply equitable enforcement principles to address network effects. Moreover, more recent legislation has integrated network economics into intellectual property law.[56]

Although patents have long protected platform technologies, such as electrical standards (e.g., AC/DC, phonogram, color television, and telecommunications),[57] the contours of intellectual property protection for network features of systems and platforms centers around software technology. Trade secrecy and contract law provided relatively effective protection for much of the software developed during the mainframe and minicomputer eras. And although advances in computer hardware fell squarely within the patent domain, there were significant doubts about the patentability of computer software into the 1990s. Hence, as microcomputers emerged, which spurred retail distribution of computer software, copyright law emerged as the primary battleground for computer software by the mid-1980s.

[55] *See* Peter S. Menell, *The Mixed Heritage of Federal Intellectual Property Law and Ramifications for Statutory Interpretation*, *in* INTELLECTUAL PROPERTY AND THE COMMON LAW 63 (Shyamkrishna Balganesh ed., 2013).

[56] *See infra* Section V.B.4 (exploring the DMCA interoperability exemption).

[57] *See* SHAPIRO & VARIAN, *supra* note 5, at 210–23.

This Part begins by discussing how trade secrecy can protect the network features of systems technologies. It then traces the evolution of copyright protection for computer software. Almost all of the major computer software battles have focused on the extent to which copyright protection afforded protection to the network features of computer software. Section V.C discusses the role of trademark and related protections for network technologies. Section V.D examines the role of patent protection for network technologies, which emerged as a more robust and controversial form of protection for computer software in the 1990s.

A. TRADE SECRET PROTECTION

Trade secret protection protects against the misappropriation of confidential information that is subject to reasonable efforts to maintain secrecy, such as security and non-disclosure agreements with employees and contractors.[58] Trade secret protection can last indefinitely, but once trade secrets become public, they lose protection.

Trade secret protection came into common usage in the software industry as a tool for protecting algorithms, software design, and coding—including APIs. Trade secret protection of passwords is also commonly used today to control access to websites and cloud servers.

Trade secret protection does not provide absolute protection for information. It only protects against misappropriation through improper means and unauthorized disclosure. Therefore, competitors do not violate trade secrecy protection through reverse engineering of publicly available products and websites. The reverse engineering limitation on trade secret protection thus exposes the trade secret owner to free riding by others. This limitation, however, strikes a salutary balance between protection on the one hand and competition and the dissemination of knowledge on the other.[59] The trade secret owner can "purchase" greater protection against this risk by investing in higher levels of security (e.g., more effective encryption for software-encoded technology). The inventor can also pursue patent protection, which proscribes reverse engineering, although only for the limited duration of the patent, and mandates disclosure of the invention to the public. By declining to pursue patent protection (or failing to satisfy the requirements

[58] *See* MENELL ET AL., *supra* note 2, at 40–152.

[59] *See* WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW (2003); Pamela Samuelson & Suzanne Scotchmer, *The Law and Economics of Reverse Engineering*, 111 YALE L.J. 1575, 1649–61 (2002); *see generally* Donald S. Chisum, Rochelle Cooper Dreyfuss, Paul Goldstein, Robert A. Gorman, Dennis S. Karjala, Edmund W. Kitch, Peter S. Menell, Leo J. Raskind, Jerome H. Reichman & Pamela Samuelson, *LaST Frontier Conference Report on Copyright Protection of Computer Software*, 30 JURIMETRICS 15, 16–18 (1989).

thereof), however, inventors should not be able to secure potentially perpetual rights in technologies merely by encrypting them or otherwise obscuring how they function. To do so would undermine the larger balance of the federal intellectual property system.

As the next Section explains, courts have interpreted copyright law to permit multiple reproductions of copyrighted software programs as a means for reverse engineering unprotected (by copyright), but secret, elements of code necessary for interoperability.

B. Copyright Protection

As the proliferation of microcomputers seeded a market for computer programs, software entrepreneurs saw copyright as an effective strategy to protect their programs from unauthorized reproduction and distribution. Computer software, however, does not fit easily within the copyright mold. Copyright law had long denied protection to functional elements. Although written in text, computer software provides the gears and levers for digital machines—which fits more naturally within the utility patent system.[60]

The rapid emergence of the computer software marketplace in the early 1970s posed a dilemma for intellectual property policymakers. Computer software could be expensive to develop and was easily pirated, creating a severe appropriability problem for the nascent software industry.[61] Patent law, which had long served as the primary form of protection for technological advances in machines and processes, was thought to be too costly, time-consuming, stringent, and uncertain as a means for protecting software products against piracy.[62] Copyright law had long provided an effective means of protecting literary works from piracy, but its doctrines excluding ideas and functional elements from protection raised serious questions about its appropriateness for protecting inherently utilitarian works. Copyright's low threshold for protection (mere originality), broad array of rights (including the right to adapt), and long duration created a high risk of overbroad protection for computer software products, in direct opposition to the parsimony principle. On the other hand, copyright law's limiting principles, such as the idea-expression dichotomy (denying copyright protection to expression that

[60] *Cf.* Baker v. Selden, 101 U.S. 99, 102 (1879) ("The claim to an invention or discovery of an art or manufacture must be subjected to the examination of the Patent Office before an exclusive right therein can be obtained; and it can only be secured by a patent from the government.").

[61] *See* Bill Gates, *An Open Letter to Hobbyists*, 2 HOMEBREW COMPUTER CLUB NEWSL. 2 (1976).

[62] *See* Menell, *supra* note 23, at 1347–51.

encumbers the use of ideas) and the fair use doctrine, provided tools for aligning copyright protection with the parsimony principle.

The interplay of copyright protection and network effects has played out on several fronts during the past four decades. Section V.B.1 explains the principal legislation undergirding copyright protection for computer software. Section V.B.2 traces the development of software copyright jurisprudence relating to APIs through 2010. Section V.B.3 explores software licensing and the emergence and growth of the free and open source movements—key drivers of network technology markets. Section V.B.4 explores the interoperability exception to the anti-circumvention provisions added to the copyright law in 1998. Section V.B.5 picks up where Section V.B.2 left off by examining the *Oracle v. Google* litigation.[63] Section V.B.6 examines copyright protection for standards and codes.

1. *Software Copyright Legislation: The Copyright Act of 1976, the CONTU Report, and the 1980 Amendments*

The software protection controversy of the early 1970s emerged at an inopportune time. Congress had been working for nearly two decades to overhaul the Copyright Act of 1909 and was nearing closure in the early to mid-1970s. Faced with the challenge of fitting computer and other new information technologies under the existing umbrella of intellectual property protection, Congress established the National Commission on New Technological Uses of Copyrighted Works (CONTU) to study the implications of the new technologies and recommend revisions to federal intellectual property law. As a stopgap, Congress included computer software within the scope of "literary works" in the Copyright Act of 1976 ("1976 Act"). The House Report explains:

> The term "literary works" does not connote any criterion of literary merit or qualitative value: it includes catalogs, directories, and similar factual, reference, or instructional works and *compilations of data*. It also includes *computer data bases*, and *computer programs* to the extent that they incorporate authorship in the programmer's expression of original ideas, as distinguished from the ideas themselves.[64]

[63] *See* Oracle Am., Inc. v. Google Inc., 750 F.3d 1339 (Fed. Cir. 2014); Oracle Am., Inc. v. Google Inc., 872 F. Supp. 2d 974 (N.D. Cal. 2012).

[64] H.R. REP. No. 94-1476, at 54 (1976) (emphasis added).

Other provisions of the 1976 Act, however, maintained traditional exclusions for ideas and functional features.[65]

The CONTU Final Report concluded that copyright law should protect the intellectual work embodied in computer software, notwithstanding the fundamental principle that copyright cannot protect "any idea, procedure, process, system, method of operation, concept, principle, or discovery" and the Supreme Court's foundational *Baker v. Selden* decision.[66] Nonetheless, CONTU recommended that Congress immunize rightful possessors of a computer program from liability for using the program (which typically results in reproduction of computer code) and making a backup copy of computer programs, which Congress largely adopted in 1980.[67]

In keeping with copyright law's fundamental limiting principles, the CONTU Final Report explained that while "one is always free to make a machine perform any conceivable process (in the absence of a patent), [] one is not free to take another's program," subject to copyright's limiting doctrines–originality and the idea-expression dichotomy.[68] The Report further explained that

> [t]he "idea-expression identity" exception provides that copyrighted language may be copied without infringing when there is but a limited number of ways to express a given idea. This rule is the logical extension of the fundamental principle that copyright cannot protect ideas. In the computer context this means that when specific instructions, even though previously copyrighted, are the only and essential means of accomplishing a given task, their later use by another will not amount to an infringement.[69]

Thus, while recognizing important limitations on copyright protection for computer software, including the § 102(b) limitations, Congress intended that software programmers would garner protection for their programming design and coding choices to the extent that the expression was separable from the

[65] 17 U.S.C. § 102(b) (2018) ("In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery.").

[66] *Id.*; *see* Baker v. Selden, 101 U.S. 99 (1879); *see also* NAT'L COMM'N ON NEW TECHNOLOGICAL USES OF COPYRIGHTED WORKS (CONTU), FINAL REPORT OF THE NATIONAL COMMISSION ON NEW TECHNOLOGICAL USES OF COPYRIGHTED WORKS 18–19 (1978) [hereinafter CONTU REPORT].

[67] 17 U.S.C. §§ 101 (adding a definition of "computer program"), 117 (providing for limitations on exclusive rights on computer programs, including the making of additional copies for archival purposes).

[68] *See* CONTU REPORT, *supra* note 66, at 20.

[69] *Id.*

underlying ideas. In this way, the general programming ideas and unoriginal programming choices remain free for others to use while the creative effort in particularized programming choices and compilations, especially in complex programs, gains protection from copyists.

2. *Software Copyright Jurisprudence: The First Wave*

The 1976 Copyright Act, as well as the CONTU Report, pushed the availability and scope of copyright protection for computer software to the courts. The treatment of APIs under copyright law emerged over the next two decades as courts interpreted and applied the § 102(b) limitations (including the idea-expression dichotomy), infringement standards, the fair use defense, and other legal doctrines and standards. Courts confronted battles across various software markets—from microcomputer operating systems to job scheduling software for mainframe computers, mobile phone networks, computer-user interfaces, video game devices, printer cartridges, garage door openers, and all manner of application programs (such as business systems, design programs, video games, and spreadsheets). Nearly every major software copyright litigation involved interoperability elements.

After an inauspicious start, the federal courts implemented a balanced framework for both protecting computer software against piracy and interpreting the idea-expression doctrine to ensure that copyright law excludes functional features of computer technology.[70] These decisions effectuated the subtle balance to which the CONTU Report referred. The courts came to appreciate that "creativity" must be understood contextually. While programming a computer can unquestionably be termed "creative" in a general sense, it is not necessarily "creative" in a copyright sense. Just as the design of an efficient mechanical machine can be creative, such devices are not eligible for copyright protection unless the aesthetic features can be separated from the functional attributes.[71] Lines of code are the gears and levers of digital machines. The fact that computer software, like a sculptural work, is eligible for copyright protection does not authorize protection for functional features.

The courts came to recognize that APIs have significant functional dimensions. They serve in many contexts as the basis for interoperability of

[70] *See generally* Peter S. Menell, *An Epitaph for Traditional Copyright Protection of Network Features of Computer Software*, 43 ANTITRUST BULL. 651, 661–72 (1998).

[71] 17 U.S.C. § 101 (" 'Pictorial, graphic, and sculptural works' include two-dimensional and three-dimensional works . . . the design of a useful article . . . shall be considered a pictorial, graphic, or sculptural work only if, and only to the extent that, such design incorporates pictorial, graphic, or sculptural features that can be identified separately from, and are capable of existing independently of, the utilitarian aspects of the article.").

computer technologies. The First Circuit held that the particular functional specifications, as opposed to the implementing code, can be fairly characterized as "methods of operation." Although the Supreme Court's split decision in *Lotus v. Borland* left some uncertainty,[72] the resolution of that litigation marked the end of the major API copyright litigations that had raged since the early 1980s.

This Section traces that evolution. Section a) examines the emergence of jurisprudence excluding functional and network features of computer software. Section b) explores the related issue of whether competitors can reproduce computer software as a means of learning unprotectable code elements.

a) Unprotectability of Functional and Network Features

The first major cases to address copyright protection for interoperable features of computer software pitted Apple Computer Corporation, then a young, break-out microcomputer company, against cavalier, unscrupulous competitors offering discount "interoperable" Apple II clones.[73] The clone makers quickly entered the market by simply copying, bit by bit, Apple's operating system and application programs.

The defendants in these cases argued that copyright protection did not extend to non-human readable (object code) formats of computer software and that the idea-expression doctrine barred copyright protection for operating system programs. They further argued that copyright protection should not stand in the way of their selling computers that can run programs written for the Apple II. The courts had little trouble validating Apple's complaint that verbatim copying of millions of bits of code constituted copyright infringement. The 1976 Act, in conjunction with the CONTU Report, clearly extended copyright protection in these circumstances.

Unfortunately, the Third Circuit's decision included language suggesting that copyright protection could encompass the functional requirements for interoperability: "total compatibility with independently developed application programs . . . is a commercial and competitive objective which does not enter

[72] *See* Lotus Dev. Corp. v. Borland Int'l, 516 U.S. 233 (1996) (affirming, without opinion by an equally divided vote, the First Circuit's decision holding that the menu command structure for a spreadsheet is an uncopyrightable method of operation under § 102(b)).

[73] *See* Apple Comput., Inc. v. Franklin Comput. Corp., 545 F. Supp. 812 (E.D. Pa. 1982), *rev'd*, 714 F.2d 1240 (3d Cir. 1983); Apple Comput., Inc. v. Formula Int'l, Inc., 562 F. Supp. 775 (C.D. Cal. 1983), *aff'd*, 725 F.2d 521 (9th Cir. 1984).

into the somewhat metaphysical issue of whether particular ideas and expressions have merged."[74] Since two entirely different programs can achieve the same "certain result[s]"[75]—for example, generate the same set of protocols needed for interoperability—the court was not justified in making such an expansive statement about the scope of copyright protection for computer program elements. CONTU was clear that "one is always free to make the machine do the same thing as it would if it had the copyrighted work placed in it, but only by one's own creative effort rather than by piracy."[76] Given the verbatim copying of millions of bits of object code, there was no need to address the interoperability issue. The defendant failed to explain which elements of the program were protectable and which were not.

The Third Circuit's decision in *Whelan Associates, Inc. v. Jaslow Dental Laboratory, Inc.*[77] further expanded copyright protection for computer software. In that case, Jaslow Dental Laboratory had hired Whelan Associates, a custom software company, to develop a computer program to organize its bookkeeping and administrative tasks. When Jaslow developed and marketed its own program for managing a dental laboratory, Whelan sued Jaslow for copyright infringement. The evidence at trial showed that although Jaslow had not literally copied Whelan's code, there were overall structural similarities between the two programs. As a means of distinguishing protectable expression from unprotectable idea, the court reasoned:

> [*T*]*he purpose or function of a utilitarian work would be the work's idea, and everything that is not necessary to that purpose or function would be part of the expression of the idea* . . . . Where there are various means of achieving the desired purpose, then the particular means chosen is not necessary to the purpose; hence, there is expression, not idea.[78]

In applying this rule, the court defined the idea as "the efficient management of a dental laboratory," which countless programs could express.[79] Drawing the idea-expression dichotomy at such a high level of abstraction implied an expansive scope of copyright protection. Although the case did not directly address copyright protection for interoperable features of computer code, the court's mode of analysis expanded the scope of copyright protection to all aspects of computer programs. If everything below the general purpose of the program were protectable under copyright law, then it would follow that

[74] Apple Comput. v. Franklin Comput. Corp., 714 F.2d 1240, 1253 (3d Cir. 1983).

[75] 17 U.S.C. § 101 (definition of "computer program").

[76] *See* CONTU REPORT, *supra* note 66, at 21.

[77] 797 F.2d 1222 (3d Cir. 1986).

[78] *Id.* at 1236 (emphasis in original) (citations omitted).

[79] *Id.* at 1236 n.28.

particular protocols were protectable because there would be other ways to accomplish the program's same general purpose. Such a result would effectively bar competitors from developing interoperable programs and computer systems.

Commentators roundly criticized the *Whelan* test,[80] and other courts developed alternative approaches. A few months after *Whelan*, the Fifth Circuit confronted a similar claim of copyright infringement based upon structural similarities between two programs designed to provide cotton growers with information regarding cotton prices and availability, accounting services, and a means for conducting cotton transactions electronically.[81] In declining to follow the *Whelan* approach, the court found that the similarities in the programs were dictated largely by standard practices in the cotton market—what the court called "externalities"—such as the "cotton recap sheet" for summarizing basic transaction information, which constitute unprotectable ideas. The court found persuasive the decision in *Synercom Technology, Inc. v. University Computing Co.*, which analogized the "input formats" of a computer program (the organization and configuration of information to be inputted into a computer) to the "figure-H" pattern of an automobile stick shift.[82]

Drawing on the Fifth Circuit's approach and Judge Learned Hand's foundational test for analyzing copyright infringement,[83] the Second Circuit crafted what has become the leading framework for analyzing infringement of computer software code.[84] Computer Associates (CA), a leading mainframe software provider, had developed a job-scheduling program (SCHEDULER) for IBM mainframe computers. Part of the success of this program was that it had a sub-component (ADAPTER) which interoperated with any of the three IBM mainframes. Thus, the user did not need to customize its programs for each of the IBM mainframes. CA's ADAPTER program ensured that programs written for SCHEDULER would run on any of the three IBM mainframes.

[80] *See* Chisum et al., *supra* note 59, at 20–21; Menell, *supra* note 51, at 1074; Note, Steven R. Englund, *Idea, Process, or Protected Expression?: Determining the Scope of Copyright Protection of the Structure of Computer Programs*, 88 MICH. L. REV. 866, 881 (1990).

[81] *See* Plains Cotton Coop. Ass'n v. Goodpasture Comput. Serv., Inc., 807 F.2d 1256 (5th Cir. 1987).

[82] 462 F. Supp. 1003, 1013 (N.D. Tex. 1978).

[83] *See* Nichols v. Universal Pictures Corp., 45 F.2d 119 (2d Cir. 1930) (espousing the idea-expression doctrine, that ideas are not copyrightable but expression of those ideas may be subject to copyright protection).

[84] *See* Comput. Assocs. Int'l v. Altai, Inc., 982 F.2d 693 (2d Cir. 1992).

CA sued Altai, a competitor that pursued a similar strategy for designing its job scheduling software for the IBM mainframes. Unbeknownst to Altai's management, one of its key programmers copied 30% of ADAPTER code into Altai's job scheduling software product. When Altai management learned of the copying, the company initiated a "clean room" process to insulate its programmers from copyright-protected code so as to ensure that the resulting program interoperated with the IBM mainframes without copying any ADAPTER code.[85]

Altai accepted responsibility for copyright infringement based on the early version. Nonetheless, drawing on the Third Circuit's *Whelan* decision, CA claimed that the clean room version was also infringing due to structural similarities at various levels, such as flow charts, inter-modular relationships, parameter lists, and macros. The Second Circuit rejected *Whelan*'s approach. As an alternative, it put forth a systematic analytical framework for determining copyright infringement of computer code:

> In ascertaining substantial similarity . . . a court would first break down the allegedly infringed program into its constituent structural parts. Then, by examining each of these parts for such things as incorporated ideas, expression that is necessarily incidental to those ideas, and elements that are taken from the public domain, a court would then be able to sift out all non-protectable material. Left with a kernel, or possibly kernels, of creative expression after following this process of elimination, the court's last step would be to compare this material with the structure of an allegedly infringing program.[86]

The court's "abstraction-filtration-comparison" test recognized that an idea could exist at multiple levels of a computer program and not solely at the most abstract level. Furthermore, the ultimate comparison is not between the programs in their entirety. Rather, courts must focus solely on whether *protectable* elements of the program were copied. Of most importance for fostering interoperability, the court held that copyright protection did not extend to those program elements where the programmer's "freedom to choose" is

> circumscribed by extrinsic considerations such as (1) the mechanical specifications of the computer on which a particular program is intended to run; (2) compatibility requirements of other programs with which a program is designed to operate in conjunction; (3) computer manufacturers' design standards; (4) demands of the

[85] P. Anthony Sammi, Christopher A. Lisy & Andrew Gish, *Good Clean Fun: Using Clean Room Procedures in Intellectual Property Litigation*, 25 INTELL. PROP. & TECH. L.J. 3, 6 (2013).

[86] *Comput. Assocs. Int'l*, 982 F.2d at 706.

> industry being serviced; and (5) widely accepted programming practices within the computer industry.[87]

Directly rejecting the Third Circuit's dictum in *Apple v. Franklin*[88] that achieving "total compatibility with independently developed application programs . . . is a commercial and competitive objective which does not enter into the somewhat metaphysical issue of whether particular ideas and expressions have merged," the Second Circuit recognized that external factors such as interface specifications, de facto industry standards, and accepted programming practices are not protectable under copyright law. The formulation of the Second Circuit test judges these external factors when the allegedly infringing activities (i.e., ex-post) occur, not when the first program is written. The court emphasized that the first company to write a program for a particular application should not be able to " 'lock up' basic programming techniques as implemented in programs to perform particular tasks."[89]

Other circuits embraced the Second Circuit's *Altai* framework.[90] The *Altai* case addressed programmers' freedom to write code to interoperate with externally established APIs—in that case by IBM. IBM had not challenged CA's or Altai's use of its interface specifications. It welcomed other companies to develop software for its mainframes. Thus, the case did not specifically address whether the API developer could assert a copyright infringement claim based on unauthorized use of their interface specifications. That issue would emerge in a series of cases involving video games and spreadsheets.

The Ninth Circuit's decision in *Sega Enterprises Ltd. v. Accolade, Inc.* expressly recognized the legitimacy of deciphering and copying particular lock-out codes for purposes of developing interoperable products.[91] Sega developed a successful video game platform (Genesis) for which it licensed access to video game developers. Accolade, a video game manufacturer, wanted to distribute versions of its game on the Genesis platform. It did not, however, want to limit distribution exclusively to Genesis, as Sega required. Rather than license access to Sega's code, Accolade reverse engineered the access code through a painstaking effort that entailed making hundreds of intermediate copies of

[87] *Id.* at 709–10.

[88] *See* Apple Comput., Inc. v. Franklin Comput. Corp., 714 F.2d 1240, 1253 (3d Cir. 1983).

[89] *See Comput. Assocs. Int'l*, 982 F.2d at 712 (quoting Menell, *supra* note 51, at 1087).

[90] *See* Gates Rubber Co. v. Bando Chem. Indus., Ltd., 9 F.3d 823, 836–43 (10th Cir. 1993); Eng'g Dynamics, Inc. v. Structural Software, Inc., 26 F.3d 1335 (5th Cir. 1994); Apple Comput., Inc. v. Microsoft Corp., 35 F.3d 1435 (9th Cir. 1994); Bateman v. Mnemonics, Inc., 79 F.3d 1532, 1547 (11th Cir. 1996); Mitel, Inc. v. Iqtel, Inc., 124 F.3d 1366 (10th Cir. 1997).

[91] 977 F.2d 1510 (9th Cir. 1993).

Sega's computer code. Accolade then incorporated only those code elements (approximately 25 bytes in games containing between 500,000 and 1.5 million bytes) that were necessary to achieve interoperability with the Genesis platform into Accolade game cartridges.

Sega sued Accolade for copyright infringement. Given the relatively small amount of Sega code in the Accolade game cartridges, Sega focused its copyright claim on Accolade's reproduction of the entirety of Sega's program code for purposes of isolating those code elements needed to interoperate with the Genesis console. The district court rejected Accolade's argument that such intermediate copies—made solely for the purpose of reverse engineering the platform—constituted fair use and granted a preliminary injunction. The Ninth Circuit held that "disassembly of object code in order to gain an understanding of the ideas and functional concepts embodied in the code is a fair use that is privileged by section 107 of the Act."[92] Balancing these factors, the Ninth Circuit ruled that "the functional requirements for compatibility with the Genesis [video game console are] aspects of Sega's programs that are not protected by copyright."[93] In effect, the court held that copyright law does not protect the particular code or process needed for interoperating with a copyrighted computer program (such as lockout code). The Ninth Circuit reaffirmed and expanded the *Sega* decision in *Sony Computer Entertainment, Inc. v. Connectix Corp.*[94]

The Northern District of California and the Ninth Circuit applied the *Altai* framework to the graphical user interface features of a computer program in *Apple Computer, Inc. v. Microsoft Corp.*[95] Apple alleged that Microsoft's Windows operating system infringed copyrights in the desktop graphical user interface of its Macintosh computer system. A licensing agreement authorizing Microsoft to use aspects of Apple's graphical user interface muddied the copyright issue. The court determined, however, that the licensing agreement was not a complete defense to the copyright claims and therefore undertook an analysis of the scope of copyright protection for a large range of audiovisual elements of computer screen displays.

In framing the analysis, the district court expressly recognized the relevance of network externalities and the cumulative nature of innovation to the scope of copyright protection:

[92] *Id.* at 1518.

[93] *Id.* at 1522 (citing 17 U.S.C. § 102(b)).

[94] 203 F.3d 596 (9th Cir. 2000).

[95] 799 F. Supp. 1006 (N.D. Cal. 1992), *aff'd in part*, *rev'd in part*, 35 F.3d 1435 (9th Cir. 1994).

Copyright's purpose is to overcome the public goods externality resulting from the non-excludability of copier/free riders who do not pay the costs of creation. Peter S. Menell, *An Analysis of the Scope of Copyright Protection for Application Programs*, 41 STAN. L. REV. 1045, 1059 (1989). But overly inclusive copyright protection can produce its own negative effects by inhibiting the adoption of compatible standards (and reducing so-called "network externalities"). Such standards in a graphical user interface would enlarge the market for computers by making it easier to learn how to use them. *Id.* at 1067-70. Striking the balance between these considerations, especially in a new and rapidly changing medium such as computer screen displays, represents a most ambitious enterprise. *Cf Lotus Dev. Corp. v. Paperback Software Int'l*, 740 F. Supp. 37 (D. Mass. 1990).

While the Macintosh interface may be the fruit of considerable effort by its designers, its success is the result of a host of factors, including the decision to use the Motorola 68000 microprocessor, the tactical decision to require uniform application interfaces, and the Macintosh's notable advertising. And even were Apple to isolate that part of its interface's success owing to its design efforts, lengthy and concerted effort alone "does not always result in inherently protectible expression." [quoting *Computer Associates v. Altai*, 982 F.2d at 711.]

By virtue of having been the first commercially successful programmer to put these generalized features together, Apple had several years of market dominance in graphical user interfaces until Microsoft introduced Windows 3.0, the first DOS-based windowing program to begin to rival the graphical capability of the Macintosh . . . . To accept Apple's "desktop metaphor"/"look and feel" arguments would allow it to sweep within its proprietary embrace not only Windows and NewWave but, at its option, also other desktop graphical user interfaces which employ the standardized features of such interfaces, and to do this without subjecting Apple's claims of copyright to the scrutiny which courts have historically employed. Apple's copyrights would hold for programs in existence now or in the future—for decades. One need not profess to know for sure where should lie the line between expression and idea, between protection and competition to sense with confidence that this would afford too much protection and yield too little competition.

The importance of such competition, and thus improvements or extensions of past expressions, should not be minimized. The Ninth Circuit has long shown concern about the uneasy balance which copyright seeks to strike: "[w]hat is basically at stake is the extent of the copyright owner's monopoly—from how large an area of activity

> did Congress intend to allow the copyright owner to exclude others?"[96]

The court found that most of the similar iconsbetween Apple's graphical user interface and Microsoft's Windows that were not authorized by the licensing agreements were either not lacking originality or subject to one or more of copyright's limiting doctrines. Drawing on the principle that compilations of largely uncopyrightable elements are only protected against "bodily appropriation of expression,"[97] the court applied a "virtual identity" standard to compare the works as a whole and determined that no infringement had occurred.[98] On appeal, the Ninth Circuit affirmed the district court's dissection of the works to determine which elements are protectable, its filtering of unprotectable elements, and its application of the "virtual identity" standard.[99]

The copyrightability of command systems for computer software arose most directly in litigation surrounding spreadsheet technology. Building upon the success of the VisiCalc program developed for the Apple II computer, Lotus Corporation marketed an enhanced operating spreadsheet program incorporating many of VisiCalc's features and commands into its 1-2-3 program for the IBM PC platform. Lotus 1-2-3 quickly became the market leader for spreadsheets running on IBM and IBM-compatible machines, and knowledge of the program became a valuable skill in the accounting and management fields. The 1-2-3 command hierarchy was particularly attractive because it logically structured more than 200 commands (see Figure 1). Users could create custom programs (called macros) to automate particular accounting and business planning tasks. Businesses and users increasingly became "locked-in" to the 1-2-3 command structure as they invested time to learn the system and their libraries of macros grew.[100] By the late 1980s, software developers seeking to enter the spreadsheet market could not ignore the large premiums that consumers placed on their investments in the 1-2-3 system.[101]

[96] *Apple Comput., Inc.*, 799 F. Supp. at 1025–26 (quoting Herbert Rosenthal Jewelry Corp. v. Kalpakian, 446 F.2d 738, 742 (9th Cir. 1971)).

[97] *See* Harper House, Inc. v. Thomas Nelson, Inc., 889 F.2d 197, 205 (9th Cir. 1989).

[98] *See* Apple Comput., Inc. v. Microsoft Corp., 35 F.3d 1435, 1446 (9th Cir. 1994) (holding that "[u]nder *Harper House* and *Frybarger [v. International Business Machines Corp.*, 812 F.2d 525 (9th Cir. 1987)], there can be no infringement unless the works are virtually identical").

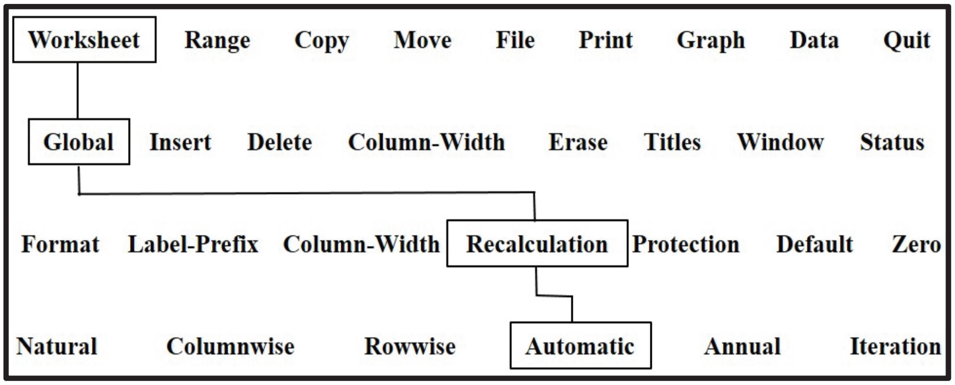
[99] *See id.* at 1446–47.

[100] *See generally* Gandal, *supra* note 4.

[101] *See* Menell, *supra* note 70, at 697.

Figure 1

Lotus 1-2-3 Menu Command Hierarchy



After three years of intensive development efforts, Borland International, developer of several successful software products including Turbo Pascal and Sidekick, introduced Quattro Pro, its entry into the spreadsheet market.[102] Quattro Pro offered improved design and graphics over Lotus 1-2-3. Computer magazines praised its innovation.[103] Quattro Pro offered a new interface for its users, which many preferred over the 1-2-3 interface. Nonetheless, because of the large number of users already familiar with the 1-2-3 command structure and those who had made substantial investments in developing 1-2-3 macros, Borland considered it essential to offer an operational mode based on the 1-2-3 command structure as well as macro compatibility. Nonetheless, Borland's visual representation of the 1-2-3 command mode substantially differed from the 1-2-3 screen displays.

Lotus sued Borland for copyright infringement based on Quattro Pro's emulation of the 1-2-3 menu command hierarchy.[104] The First Circuit viewed the case as one of first impression: "[w]hether a computer menu command hierarchy constitutes copyrightable subject matter."[105] The court distinguished *Altai* as dealing with protection of computer code as opposed to the results of such code. Instead, the First Circuit saw the subject matter of the *Lotus* case as a "method of operation" falling directly within the exclusions from copyright

[102] *See generally* Peter S. Menell, *Envisioning Copyright Law's Digital Future*, 46 N.Y.L. SCH. L. REV. 63, 91–93 (2003) (providing background on Borland and the *Lotus v. Borland* litigation).

[103] *See* Peter S. Menell, *An Epitaph for Traditional Copyright Protection of Network Features of Computer Software*, 43 ANTITRUST BULL. 651, 698 n.137 (1998).

[104] *See* Lotus Dev. Corp. v. Borland Int'l, Inc., 799 F. Supp. 203 (D. Mass. 1992), *rev'd*, 49 F.3d 807, 813 (1st Cir. 1995).

[105] Lotus Dev. Corp. v. Borland Int'l, Inc., 49 F.3d 807, 813 (1st Cir. 1995), *aff'd by equally divided Court*, 516 U.S. 233 (1996).

protection set forth in 17 U.S.C. § 102(b). The court held the Lotus menu command hierarchy is an uncopyrightable "method of operation."

> The Lotus menu command hierarchy provides the means by which users control and operate Lotus 1-2-3. If users wish to copy material, for example, they use the "Copy" command. If users wish to print material, they use the "Print" command. Users must use the command terms to tell the computer what to do. Without the menu command hierarchy, users would not be able to access and control, or indeed make use of, Lotus 1-2-3's functional capabilities.
>
> The Lotus menu command hierarchy does not merely explain and present Lotus 1-2-3's functional capabilities to the user; it also serves as the method by which the program is operated and controlled.[106]

The U.S. Supreme Court affirmed without opinion by an equally divided vote.[107]

Subsequent appellate decisions reached similar outcomes, although they did not fully adopt the First Circuit's categorical exclusion of menu command hierarchies from copyright protection. In *MiTek Holdings, Inc. v. ARCE Engineering Co.*,[108] the holder of a copyright in an application program that designed and arranged wood trusses for the framing of building roofs brought an infringement action against the maker of a competing program that featured a similar menu command tree and user interface. Affirming the lower court's decision, the Eleventh Circuit held that the menu and submenu command structure of the truss design program was uncopyrightable under § 102(b) of the Copyright Act because it represents a process.[109] The court did not need to reach the broader question, addressed in *Lotus*, of whether all menu command structures are uncopyrightable as a matter of law.[110]

In *Mitel, Inc. v. Iqtel, Inc.*,[111] Mitel, the maker of a widely-adopted computer system for automating the selection of a particular telephone long distance carrier and remotely activating optional telecommunications features such as speed dialing, sued Iqtel, a competing firm that used the identical command codes for copyright infringement. Because Mitel's system had become a de facto standard, Iqtel defended its use of compatible controller codes on the ground that "technicians who install call controllers would be unwilling to

[106] *Lotus*, 49 F.3d at 815.

[107] *See* Lotus Dev. Corp. v. Borland Int'l, Inc., 516 U.S. 233 (1996).

[108] 89 F.3d 1548 (11th Cir. 1996).

[109] *See id.* at 1556–57.

[110] *See id.* at 1557.

[111] 124 F.3d 1366 (10th Cir. 1997).

learn Iqtel's new set of instructions in addition to the Mitel command code set, and the technicians' employers would be unwilling to bear the cost of additional training."[112]

As Borland had done, Iqtel's product included both its own set of command codes as well as a "Mitel Translation Mode."[113] While commenting that a method of operation may in some circumstances contain copyrightable expression, the Tenth Circuit nonetheless concluded that the Mitel command codes, which were arbitrarily assigned, lacked the minimal degree of creativity necessary to qualify for copyright protection.[114] The court further held that Mitel's command codes should be denied copyright protection under the *scènes à faire* doctrine because external factors, such as compatibility requirements and industry practices, largely dictated the codes.[115]

There were no further cases reported addressing copyright protection for APIs over the next fifteen years. We address the Federal Circuit's decision upholding copyright protection for APIs in the *Oracle v. Google* case in Section V.B.5.

b) Permissibility of Reverse Engineering

As discussed in Section V.A, network system developers can use encryption and trade secret law to protect computer code.[116] Distributing computer programs in object code (binary) format typically constitutes a reasonable effort to maintain secrecy. As noted, however, competitors can lawfully gain access to such information through reverse engineering. One such method is to experiment with object code to determine which bits are necessary for interoperability. Such forensic work typically requires the investigator to make many copies, raising the risk of copyright infringement.

The LaST Frontier Final Report, a consensus statement of leading intellectual property scholars, opined that "limited copying of programs for the purpose of examination and study . . . falls within the rigorous terms of the fair use provisions in section 107 of the Copyright Act."[117] In addition to holding that computer code necessary for interoperability is unprotectable under § 102(b), the Ninth Circuit's *Sega* decision authorized the copying of

[112] *Id.* at 1369.

[113] *See id.* at 1368–70.

[114] *See id.* at 1372–74.

[115] *See id.* at 1374–76.

[116] *See supra* Section V.A.

[117] *See* Chisum et al., *supra* note 59, at 25; *see also* Samuelson & Scotchmer, *supra* note 59, at 1650.

entire computer programs for purposes of deciphering unprotectable code elements.[118] In explaining why disassembly and reproduction of object code constitute fair use, the court reasoned that the "functional requirements for compatibility" with a computer program are unprotectable by copyright.[119] The Ninth Circuit based its analysis on the architecture of the intellectual property system:

> [D]isassembly of the object code in Sega's video game cartridges was necessary in order to understand the functional requirements for Genesis compatibility. The interface procedures for the Genesis console are distributed for public use only in object code form, and are not visible to the user during operation of the video game program. Because object code cannot be read by humans, it must be disassembled, either by hand or by machine . . . If disassembly of copyrighted object code is *per se* an unfair use, the owner of the copyright gains a *de facto* monopoly over the functional aspects of his work—aspects that were expressly denied copyright protection by Congress. 17 U.S.C. § 102(b). In order to enjoy a lawful monopoly over the idea or functional principle underlying a work, the creator of the work must satisfy the more stringent standards imposed by the patent laws. *Bonito Boats, Inc. v. Thunder Craft Boats, Inc.*, 489 U.S. 141, 159–64 (1989). Sega does not hold a patent on the Genesis console.[120]

The Ninth Circuit reaffirmed and expanded the *Sega* analysis in *Sony Computer Entertainment, Inc. v. Connectix Corp.*[121]

3. *Software Licensing*

Copyright law grants authors exclusive rights to copy, adapt, distribute, publicly perform, and publicly display protected works, subject to various limitations. The early computer industry, however, did not rely on proprietary control over their customers' use or adaptation of their software programs. Nor did companies restrict customers' access to source code. Rather, the industry—led by IBM and followed by Burroughs, UNIVAC, NCR, Control Data, General Electric, and RCA (often referred to as the "Seven Dwarfs" due to IBM's dominance in the computer industry)—bundled software with their

[118] *See* Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1520–27 (9th Cir. 1993); Menell, *supra* note 16, at 332–34.

[119] *See id.* at 1522 (citing 17 U.S.C. § 102(b)).

[120] Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1526 (9th Cir. 1993).

[121] 203 F.3d 596 (9th Cir. 2000). The *Sony* case held that the fair use defense applied even in a case that allowed consumers to bypass purchasing the Sony PlayStation. In *Sega*, the reverse engineered products produced by Accolade could only be run on the Sega Genesis console.

mainframes and derived revenues from leasing computer usage and sales of complementary products and services.[122] In this era, IBM actively facilitated sharing of software among its users as a way of increasing usage of its computers.

The structure of the computer industry and copyright's role dramatically changed during the 1970s. With technological advances creating a mini-computer market and IBM's 1969 decision to unbundle software from mainframe leasing in the face of antitrust charges, computer hardware vendors and independent software developers came to use copyright licenses to protect computer programs. The opening of a competitive proprietary software marketplace ended an era in which software was freely shared.[123]

This shift produced a backlash within the programmer community that continues to reverberate throughout the computer hardware and software industries. The rapid rise of a robust microcomputer industry followed by the creation of the Internet generated a robust, independent software marketplace. These technologies had strong and complex network effects, which have been substantially affected by software licensing practices. While many hardware and software enterprises continue to rely heavily on proprietary software licensing agreements, the programmers' backlash against restrictive software licensing as well as business strategies aimed at disrupting proprietary standards have dramatically reshaped software licensing institutions, practices, and patterns.

This Section explores this evolving landscape. Section a) traces the emergence of the free software movement, which resourcefully uses copyright licensing to promote open platforms. The movement's innovative licensing framework produced a form of network effects. Section b) examines the open source movement, based on a more permissive licensing model, which broadened the shift away from proprietary software licensing. Section c) discusses the use of dedication of software copyrights to the public domain as a third alternative for promoting network effects. Section d) surveys federal copyright preemption of licensing restrictions.

a) The Free Software Movement (General Public License)

Many independent and academic programmers, who had long enjoyed free access to source code, viewed the shift to proprietary software licensing as a debilitating restriction on collaborative research, programming freedom, and software innovation. Beginning in the early 1980s, Richard Stallman, then a

[122] *See* PAUL E. CERUZZI, A HISTORY OF MODERN COMPUTING ch. 5 (2d ed. 2003).

[123] *See* DOUGLAS E. PHILLIPS, THE SOFTWARE LICENSE UNVEILED: HOW LEGISLATION BY LICENSE CONTROLS SOFTWARE ACCESS 113–15 (2009).

researcher in MIT's Artificial Intelligence Laboratory, began a grass-roots "free software" movement. Although Stallman was vehemently opposed to intellectual property protection for computer software, he came to see that the same copyright protections that exclude competitors could be deployed to prohibit restrictions on adaptation and reuse of code and to foster open platforms.[124]

Stallman established the Free Software Foundation (FSF) in 1985 to promote users' rights to use, study, copy, modify, and redistribute computer programs. The FSF devised the General Public License (GPL) to prevent programmers from building proprietary limitations into software. The GPL guarantees end users the freedoms to run, study, share (copy), and modify the software so long as the users permit use of any derivative works on the same terms.[125] In this way, GPL software "infects" derivative works with user rights and virally spreads these rights through the collaborative software ecosystem.

Stallman targeted the development of a viable UNIX-compatible open source operating as FSF's initial goal.[126] The UNIX operating system, developed by researchers at MIT, AT&T's Bell Labs, and General Electric in the late 1960s and early 1970s offered innovative time-sharing capability.[127] It became a foundation for modern computer operating system design.[128] In 1972, two Bell Labs researchers—Dennis Ritchie, inventor of the C programming language, and Ken Thompson—rewrote UNIX in C, enabling UNIX to be installed on any advanced computer system. AT&T held the copyright to UNIX, which restricted its use and adaptation. Stallman sought to liberate UNIX through the GNU ("GNU's Not Unix") GPL independent re-implementation project.

Many programmers throughout the world contributed to this effort on a voluntary basis, and by the late 1980s most of the components had been assembled. The project reached fruition in 1991 when Linus Torvalds developed a UNIX-compatible kernel—the central core of the operating

[124] *See* STEVEN WEBER, THE SUCCESS OF OPEN SOURCE 47–49 (2004).

[125] *See* Carver, *supra* note 15, at 443–44.

[126] *See Initial Announcement*, GNU OPERATING SYS. (Sept. 7, 1983), https://www.gnu.org/gnu/initial-announcement.en.html [https://perma.cc/B4X8-UNCP]; Richard Stallman, *The GNU Manifesto*, 10 DR. DOBB'S J. SOFTWARE TOOLS 30 (1985); *GNU Manifesto*, WIKIPEDIA, https://en.wikipedia.org/wiki/GNU_Manifesto [https://perma.cc/EF6V-GB84].

[127] *See Unix*, WIKIPEDIA, https://en.wikipedia.org/wiki/Unix [https://perma.cc/W7AG-GHKQ].

[128] Marshall Kirk McKusick, *Twenty Years of Berkeley Unix: From AT&T Owned to Freely Redistributable*, *in* OPEN SOURCES: VOICES FROM THE OPEN SOURCE REVOLUTION 31 (Chris DiBona, Sam Ockman & Mark Stone eds., 1999).

system.[129] Torvalds structured the evolution of his component on the GPL model. The resulting UNIX-compatible free software program, dubbed "Linux," has become widely used throughout the computing world.[130]

While attractive to many independent, non-commercial programmers, the so-called "copyleft" GPL licensing model posed a serious problem for many commercial software vendors. Although it afforded free access to GPL software, it prevented these cumulative developers from charging a royalty for their modifications and subjected further modifications by licensees to GPL restrictions.[131]

b) The Open Software Movement (Permissive Licenses)

The "open source" movement emerged as a middle ground between proprietary software distribution and the "free" software movement. Like Linux, the open source movement traces its roots to efforts to liberate UNIX. In the mid-1970s, Ken Thompson at the University of California, Berkeley, spearheaded an effort by Berkeley faculty and students to enhance UNIX capabilities.[132] In contrast to the GPL, the Berkeley Software Development (BSD) project offered its software on a "permissive" basis: licensees could distribute modifications of the BSD software whether or not the modifications were freely licensed.[133] Nonetheless, the licensee was still obliged to obtain a license from AT&T for the underlying UNIX code.[134]

As the Internet took off in the late 1990s, a growing number of hardware and software vendors embraced "free" and "open source" development and distribution strategies. They saw these non- or less-proprietary licensing models as means to prevent Microsoft from expanding its influence into the Internet and other platform technologies while simultaneously promoting competition and innovation.[135] There is now a wide variety of permissive open

[129] *See Linux*, WIKIPEDIA, https://en.wikipedia.org/wiki/Linux [https://perma.cc/VQL2-S4D8].

[130] *See id.*

[131] *See* Lothar Determann, *Dangerous Liaisons–Software Combinations as Derivative Works? Distribution, Installation, and Execution of Linked Programs under Copyright Law, Commercial Licenses, and the GPL*, 21 BERKELEY TECH. L.J. 1421, 1484 (2006).

[132] *See Berkeley Software Distribution*, WIKIPEDIA, https://en.wikipedia.org/wiki/Berkeley_Software_Distribution [https://perma.cc/SV4M-N9KN].

[133] *See BSD Licenses*, WIKIPEDIA, https://en.wikipedia.org/wiki/BSD Licenses [https://en.wikipedia.org/wiki/BSD Licenses].

[134] *See id.*

[135] *See, e.g.*, Josh Lerner & Jean Tirole, *The Economics of Technology Sharing: Open Source and Beyond*, 19 J. ECON. PERSP. 99, 106 (2005); Merges, *supra* note 39, at 191–93; Yochai Benkler, *Coase's Penguin, or, Linux and the Nature of the Firm*, 112 YALE L.J. 369, 445 (2002); Yochai

source licensing models.[136] Free (GPL) and open source software play strong and increasing roles in network technologies, such as operating systems (e.g., Linux), Internet infrastructure (e.g., Apache Web Server), and mobile devices (e.g., Android), but have been less successful in penetrating consumer as opposed to programmer-centric product areas. [137] Notwithstanding the proliferation of free and open source licenses, there have been relatively few litigated disputes.[138]

c) Dedication to the Public Domain

A further distribution alternative that has been especially important in the proliferation of network benefits is outright dedication of computer software copyrights (and other forms of intellectual property) to the public domain. Tim Berners-Lee, the developer of the World Wide Web (WWW), was initially attracted to releasing his hypertext software platform under the GPL. [139] Internet engineers, however, raised the concern that any restrictions attached to its usage could limit its adoption and use. Some large companies were rumored to be opposed to allowing usage of any software that could trigger license restrictions, including GPL copyleft requirements. Berners-Lee ultimately chose to dedicate the WWW to the public domain. Notwithstanding concerns that unprotected software could be fragmented and captured through proprietary extensions, the WWW has thrived and remained remarkably stable.[140] This is attributable to the very strong network effects of Internet protocols and the community and technically driven, open, standard-setting processes administered by the WWW Consortium (W3C) headed by Berners-Lee and the Internet Engineering Task Force (IETF).

d) Federal Preemption of Contractual Restrictions

In contrast to the free and open software movements, some software developers use licensing provisions to restrict use of their copyrighted software. Some licenses, for example, bar reverse engineering of software programs. Such a restriction affords the copyright owner greater control over

Benkler, *Sharing Nicely: On Shareable Goods and the Emergence of Sharing as a Modality of Economic Production*, 114 YALE L.J. 273, 289–91 (2004); *see generally* David McGowan, *Innovation, Uncertainty, and Stability in Antitrust Law*, 16 BERKELEY TECH. L.J. 729 (2001).

[136] MEEKER, *supra* note 12, at 287–94 (Appendix B).

[137] PHILLIPS, *supra* note 123, at 156, 158–68; *see also* Josh Lerner & Jean Tirole, *The Economics of Technology Sharing: Open Source and Beyond*, 19 J. ECON. PERSP. 99, 107 (2005).

[138] *See* Jacobsen v. Katzer, 535 F.3d 1373 (Fed. Cir. 2008); PHILLIPS, *supra* note 123, at 120–21 (2009).

[139] TIM BERNERS-LEE, WEAVING THE WEB: THE ORIGINAL DESIGN AND ULTIMATE DESTINY OF THE WORLD WIDE WEB 72–73 (2000).

[140] *See* PHILLIPS, *supra* note 123, at 174.

the development of interoperable products. The courts are divided, however, over whether federal copyright law and intellectual property policies preempt such state law, contractual provisions.

In *Vault Corp. v. Quaid Software, Ltd.*,[141] the Fifth Circuit Court of Appeals held that the Louisiana Software License Enforcement Act clause permitting a copyright owner to prohibit software decompilation or disassembly was preempted by the Copyright Act, and therefore unenforceable. A more recent case interpreted the scope of federal copyright protection more narrowly, enforcing licensing restrictions that bar activities that would otherwise fall within copyright's fair use privilege.[142] The dissenting opinion in that case, however, indicates that the scope of federal preemption of licensing restrictions that contract around the fair use privilege remains unsettled.[143] Section VI examines the related questions of whether antitrust law or misuse doctrines further restrict licensing provisions that leverage intellectual property rights to hinder downstream innovation or competition.

4. Interoperability Exception to the DMCA's Anti-Circumvention Prohibition

The permissibility of reverse engineering software to achieve interoperability arose during the legislative deliberations over the enactment of anti-circumvention prohibitions. With the emergence of the Internet in the mid-1990s, motion picture studios, record labels, publishers, and other content owners came to see encryption and other digital rights management technologies as a promising self-help means to discourage unauthorized distribution of their works. They recognized, however, that such technologies would be vulnerable to unauthorized circumvention of technological protection measures. Thus, they sought to expand copyright protection beyond its traditional prohibitions against infringement to include limits on the decrypting or circumventing of technological protection systems and the trafficking in such decryption tools. They contended that without such protection, they would be unwilling to release content onto the Internet, which in turn would hamper the adoption of broadband services. Various other interests—ranging from consumer electronics manufacturers, library associations, computer scientists, and law professors—expressed concern about potential chilling effects of such an expansion of copyright law upon those who wish to make fair use of copyrighted works.

[141] 847 F.2d 255 (5th Cir. 1988).

[142] *See* Bowers v. Baystate Techs., 320 F.3d 1317 (Fed. Cir. 2003).

[143] JONATHAN BAND & MASANOBU KATOH, INTERFACES ON TRIAL 2.0 121–33 (2011).

Congress crafted a compromise in the Digital Millennium Copyright Act of 1998 (DMCA).[144] Section 1201(a) bans circumvention of technological protection measures put in place by copyright owners to protect copyrighted works. Section (b) prohibits trafficking in anti-circumvention tools. Section 1201(f)(1) provides that

> a person who has lawfully obtained the right to use a copy of a computer program may circumvent a technological measure that effectively controls access to a particular portion of that program for the sole purpose of identifying and analyzing those elements of the program that are necessary to achieve interoperability of an independently created computer program with other programs, and that have not previously been readily available to the person engaging in the circumvention, to the extent any such acts of identification and analysis do not constitute infringement under this title.

The legislative history notes that this provision is

> intended to allow legitimate software developers to continue engaging in certain activities for the purpose of achieving interoperability to the extent permitted by law prior to the enactment of this chapter. The objective is to ensure that the effect of current case law interpreting the Copyright Act is not changed by enactment of this legislation for certain acts of identification and analysis done in respect of computer programs. *See Sega Enterprises Ltd. v Accolade, Inc.*, 977 F.2d 1510, 24 U.S.P.Q.2d 1561 (9th Cir. 1992). The purpose of this section is to foster competition and innovation in the computer and software industry.[145]

Because violations of the DMCA are not acts of copyright infringement, but rather separate offenses, courts have held that the defenses available under the Copyright Act, including fair use, do not apply to anti-circumvention violations.[146] While § 1201(c)(1) provides that "nothing in this law" shall interfere with "fair use" among other defenses, the courts have reasoned that the DMCA does not interfere with fair use but merely renders it irrelevant by allowing copyright owners to bring a non-copyright claim. Furthermore, the larger structure of the DMCA provides additional safeguards to address free expression and other concerns.

[144] 17 U.S.C. § 1201 (2018) (Circumvention of copyright protection systems).

[145] S. REP. NO. 105-190, at 13 (1998).

[146] *See* Universal City Studios, Inc. v. Corley, 273 F.3d 429 (2d Cir. 2001); 321 Studios v. MGM Studios, Inc., 307 F. Supp. 2d 1085 (N.D. Cal. 2004).

Beyond the statutory exemptions to the anti-circumvention ban, the DMCA established a triennial rulemaking process for exempting particular categories of works from the anti-circumvention ban for which "noninfringing uses by persons who are users of a copyrighted work are, or are likely to be, adversely affected."[147] Several of the granted exemptions authorize decryption for purposes of developing interoperable products.

Smartphones, tablets, other mobile computing devices, and smart TVs, all of which have networking aspects, have attracted particular attention. Several major manufacturers of these products have sought to use encryption technologies to bundle the devices in telecommunications service plans. In a series of rulemaking proceedings, the Copyright Office has exempted unlocking or "jailbreaking" of these products from the anti-circumvention ban.[148] Congress and the FCC have reinforced, extended, and expanded these exemptions.[149]

The DMCA's anti-circumvention provisions have generated several cases involving the use of technological protection measures to exclude competitors from aftermarkets—goods or services supplied for a durable product after its initial sale (e.g., replacement ink for printers). Several companies embedded digital code into their products and aftermarket components that must interoperate to function as a means of exerting control over such aftermarkets. When competitors in these aftermarkets decrypted such digital codes to manufacture their own components, these durable product manufacturers sued, alleging violation of the anticircumvention provisions of the DMCA. Some courts have declined to find liability, emphasizing that the careful balance that Congress sought to achieve between the interests of content creators and information users would be upset if the anti-circumvention prohibitions could be applied to activities that did not facilitate copyright infringement.[150]

[147] 17 U.S.C. § 1201(a)(1)(B)–(D).

[148] *See* Library of Congress, U.S. Copyright Office, Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies, 83 Fed. Reg. 54010 (Oct. 26, 2018).

[149] *See* Unlocking Consumer Choice and Wireless Competition Act, Pub. L. No. 113-144, 128 Stat. 1751 (2014); *see generally* JONATHAN BAND, INTERFACES ON TRIAL 3.0: ORACLE AMERICA V. GOOGLE AND BEYOND 79–109 (2018) (unpublished manuscript), https://ssrn.com/abstract=2876853 [https://perma.cc/NKH6-KW4S].

[150] *See* Chamberlain Group, Inc. v. Skylink Techs., Inc., 381 F.3d 1178, 1202 (Fed. Cir. 2004) (holding that section 1201 "prohibits only forms of access that bear a reasonable relationship to the protections that the Copyright Act otherwise affords copyright owners"); Lexmark Int'l, Inc. v. Static Control Components, Inc., 387 F.3d 522 (6th Cir. 2005) (holding

a) GPL 3.0 - DRM Provision

To bar intellectual property restrictions on software use and promote sharing of code, the Free Software Foundation added a provision to the General Public License 3.0 (released in 2007) barring licensors and those who use the licensed code from enforcing anti-circumvention prohibitions.[151] GPL 3.0 has not been as widely adopted as prior GPL versions, particularly among commercial enterprises.[152]

5. *Software Copyright Jurisprudence: The Oracle v. Google Litigation*

After the *Lotus v. Borland* case resolved, litigation subsided over copyright protection for the functional specifications of APIs and other network features of computer software.[153] The *Sega*, *Altai*, and *Borland* decisions and software industry norms accorded competitors the ability to develop interoperable code and devices so long as they independently implemented the functional specifications of the target platform.[154] If the programs were encrypted or only released in object code form, the competitor would need to reverse engineer the code, which could be costly and time-consuming. Beyond the drudgery of reverse engineering, copyright did not stand in the way of developing and distributing interoperable code and devices.

A shift in business strategy in the Internet Age reinforced these legal principles and industry norms. Whereas most software vendors in the pre-Internet era sought to appropriate a return on their investments directly through software and device sales and licenses, the Internet expanded the potential for multi-sided markets and indirect appropriability—principally through advertising, service plans, and use of customer data.[155] These strategies harnessed the positive feedback effects of network technologies.

that the lock-out technology at issue did not effectively control access to a copyrighted work); Storage Tech. Corp. v. Custom Hardware Eng'g & Consulting, Inc., 421 F.3d 1307 (Fed. Cir. 2005) (holding that decryption by a third party software repair entity to perform software maintenance activities is not actionable); *but see* MDY Indus., LLC v. Blizzard Entm't, Inc., 629 F.3d 928, 948–52 (9th Cir. 2010) (declining to adopt an infringement nexus requirement).

[151] *GNU General Public License: Version 3*, FREE SOFTWARE FOUND. (June 29, 2007), https://www.gnu.org/licenses/gpl-3.0.en.html [https://perma.cc/5F26-E34R] [hereinafter *GNU General Public License*].

[152] *See infra* Section V.D.3.c); *see generally* MEEKER, *supra* note 12, at ch. 10.

[153] *See* Menell, *supra* note 70, at 709–10.

[154] *See* Menell, *supra* note 16, at 448–49.

[155] *See* Martin Campbell-Kelly et al., *Economic and Business Perspectives on Smartphones as Multi-Sided Platforms*, 39 TELECOMM. POL'Y 717, 730–32 (2015); *see generally* David S. Evans, *Antitrust Economics of Multi-Sided Platform Markets*, 20 YALE J. ON REG. 325 (2003); SHAPIRO & VARIAN, *supra* note 5.

Beginning with Netscape, a growing number of Internet Age entrepreneurs valued adoptions over revenues in the start-up phase of their enterprises. The Internet provided a low-cost means of distributing information and software, goods that had zero marginal reproduction cost. For example, Sun Microsystems released the Java programming language to the public as a means of promoting its hardware sales and forestalling Microsoft's dominance of website development tools.[156] Google developed a robust revenue stream for its search technologies without ever charging users. It profited handsomely from bundling search results with keyword-generated advertisements.

Thus, many software and Internet companies welcomed adoption of their platforms, including interoperability with their APIs. Sun Microsystems dedicated the Java programing language to the public domain early on, and in 2006 licensed the Java Standard Edition, Enterprise Edition, and Micro Edition platforms—comprising packages of pre-written APIs—under the GPL. Unlike Sega, it published its API specifications for the world to see, adopt, and emulate. Its primary concern was maintaining the Write Once, Run Anywhere (WORA) interoperability of these platforms. Hence, it required licensees to verify that implementations satisfied the particular Java Technology Compatibility Kit (TCK) test.

When Google ventured into mobile platform development, it sought to take advantage of the millions of programmers intimately familiar with Java, the most widely used programming language and platform for web development. But unlike Borland, which sought to achieve perfect interoperability with the Lotus 1-2-3 menu command hierarchy so that Lotus macros could run on Borland's Quattro system, Google sought to customize Java for the smaller chip size of mobile handsets and add additional features, such as location tools and a camera. Consequently, Google did not plan to include all the Java APIs, which meant that the resulting system would not pass the Java TCK test. Moreover, Google and its open handset alliance partners did not believe that the GPL would provide sufficient flexibility for the range of players it believed would be needed to establish a robust new mobile platform. They worried that the viral share and share alike provision would discourage Google's handset manufacturer and telecommunications partners from investing in innovative features. The members of the Android Open Handset Alliance believed that a more permissive licensing model, in which

[156] *See* Menell, *supra* note 16, at 350–51.

downstream suppliers could make proprietary extensions on top of the base platform, would better promote robust competition and innovation.[157]

When licensing negotiations between Google and Sun reached an impasse, Google chose to re-implement a subset of Java API packages independently to take advantage of the vast Java programming community and the decade of testing that the Java APIs had undergone. Google did not need to reverse engineer the Java API functional specifications because Sun disclosed them. Nonetheless, Google had to devote substantial resources to re-implementing the code using a clean room process.

When Google introduced Android in late 2007, Sun's CEO publicly praised the adoption of Java. Privately, however, he and other Sun leaders seethed at Google's cavalier approach and forking of the Java platform. Nonetheless, Sun refrained from blocking Android through legal action.[158]

With its hardware business in decline and unable to monetize Java, Sun's viability as an independent company came into question. Oracle Corporation, which had built many of its software products on the Java platform, acquired Sun in 2010. Oracle immediately pressured Google to license Java and when Google declined, Oracle sued alleging that Android infringed Java-related patents and copyrights. Oracle focused its copyright claim on Google's copying of function labels, functional specifications (declarations), and the structure, sequence, and organization of 37 Java API packages.

After the jury rejected Oracle's patent causes of action, the district court ruled that the Java APIs were not copyrightable.[159] Judge Alsup cautioned that the ruling did not hold "Java API packages are free for all to use without license" or that "the structure, sequence and organization of all computer programs may be stolen."[160] He grounded his decision in the particular and distinctive functional attributes of the 37 Java APIs and that Google independently wrote its own implementing code using a clean room process.[161] The principal copying concerned the lines of declarations, which are necessary to operate the particular methods of the APIs. As Judge Alsup explained:

> Significantly, the rules of Java dictate the precise form of certain necessary lines of code called declarations, whose precise and necessary form explains why Android and Java *must be* identical when it comes to those particular lines of code. That is, since there is only

[157] *See* Menell, *supra* note 16, at 359–72.

[158] *Id.* at 369.

[159] *See* Oracle Am., Inc. v. Google Inc., 872 F. Supp. 2d 974 (N.D. Cal. 2012).

[160] *Id.* at 1002.

[161] *See id.*

> one way to declare a given method functionality, everyone using that function must write that specific line of code in the same way.[162]

While acknowledging that the overall structure of the Java API packages is creative, original, and "resembles a taxonomy," Judge Alsup nonetheless concluded that it functions as "a command structure, a system or method of operation—a long hierarchy of over six thousand commands to carry out pre-assigned functions."[163] Applying copyright's limiting doctrines as the Ninth Circuit has interpreted them, emphasizing the *Sega* decision, and following CONTU's guidance that when specific computer instructions, "*even though previously copyrighted, are the only and essential means of accomplishing a given task, their later use by another will not amount to an infringement*,"[164] Judge Alsup determined that Google was free to write code that accomplished the same functionality as the Java APIs at issue even if it did not achieve complete compatibility with the full Java platform. Later developers can achieve the *particular functionality* or method of operation of an API subsystem (and even groups of subsystems) so long as they write their own code and no patent protects that method.

Oracle appealed the copyright issues to the U.S. Court of Appeals for the Federal Circuit.[165] The Federal Circuit is bound by regional circuit law when reviewing questions that involve law and precedent not exclusively assigned to the Federal Circuit.

Notwithstanding the Ninth Circuit's holding in *Sega* and *Sony Computer Entertainment, Inc. v. Connectix Corp.* that copyright law does not prohibit the precise coding necessary to achieve interoperability,[166] the Federal Circuit reversed the district court's determination that the structure, sequence, and organization of the 37 Java APIs were not copyrightable.[167] The appellate court determined that even high-level API design choices—including function

[162] *Id.* at 979 (emphasis in original).

[163] *Id.* at 999–1000.

[164] *Id.* at 986 (quoting CONTU REPORT, *supra* note 66, at 20) (emphasis added by Judge Alsup).

[165] *See generally* Peter S. Menell, *API Copyrightability Bleak House: Unraveling and Repairing the Oracle v. Google Jurisdictional Mess*, 31 BERKELEY TECH. L.J. 1515, 1581–83 (2016) (explaining and questioning the Federal Circuit's jurisdiction over appeals from district court cases involving patent infringement allegations even if neither party challenges the district court's patent rulings).

[166] Sega Enters. v. Accolade, Inc., 977 F.2d 1510, 1525 (9th Cir. 1993); Sony Comput. Entm't., Inc., v. Connectix Corp., 203 F.3d 596, 603 (9th Cir. 2000) ("There is no question that the Sony BIOS contains unprotected functional elements.").

[167] *See* Oracle Am., Inc. v. Google, Inc., 750 F.3d 1339 (Fed. Cir. 2014); Menell, *supra* note 16, at 388.

labeling choices and compilation of functions—satisfy copyright law's low originality threshold.[168] The court side-stepped the *Sega* and *Sony* cases by construing Ninth Circuit law to hold that "copyrightability is focused on the choices available to the plaintiff at the time the computer program was created," not the defendant's desire to achieve interoperability.[169] The court concluded that Google's interoperability argument comes into play only as part of a fair use defense, an issue on which the jury had hung.[170] Consequently, the court remanded the case for a fair use trial.[171]

On remand, the jury concluded that Android's use of Java API declarations and structure, sequence, and organization constituted fair use. The Federal Circuit once again reversed, holding that the fair use balance tilted in Oracle's favor.[172] The Federal Circuit's decision gives no weight to the second fair use factor based on a questionable reading of Ninth Circuit jurisprudence.[173]

The Federal Circuit's decision rejecting Judge Alsup's API copyrightability ruling is the most significant recent federal appellate decision to confront the copyrightability of APIs. Given the proliferation of software patents, there is a high likelihood that a company with a widely-used set of APIs would be able to pursue both patent and copyright causes of action in the same litigation, thereby bringing the Federal Circuit's exclusive jurisdiction over patent cases

[168] *See Oracle*, 750 F.3d at 1354, 1356–57.

[169] *See id.* at 1370–71.

[170] *See id.* at 1358 (citing Ets-Hokin v. Skyy Spirits, Inc., 225 F.3d 1068, 1082 (9th Cir. 2000)); Satava v. Lowry, 323 F.3d 805, 810 n.3 (9th Cir. 2003) ("The Ninth Circuit treats scènes à faire as a defense to infringement rather than as a barrier to copyrightability.").

[171] *See id.* at 1372–74.

[172] *See* Oracle America, Inc. v. Google LLC, 886 F.3d 1179 (Fed. Cir. 2018).

[173] *See id.* at 1205 (explaining that:

> [t]he Ninth Circuit has recognized . . . that th[e] second factor 'typically has not been terribly significant in the overall fair use balancing.' Dr. Seuss Enters., L.P. v. Penguin Books USA, Inc., 109 F.3d 1394, 1402 (9th Cir. 1997) (finding that the 'creativity, imagination and originality embodied in The Cat in the Hat and its central character tilts the scale against fair use'); Mattel[, Inc. v. Walking Mountain Prods., 353 F.3d 792, 803 (9th Cir. 2003)] (similar).

The Federal Circuit's reliance on *Dr. Seuss Enters.* and *Mattel* is misplaced. Those cases addressed familiar children's stories and dolls; neither involved functional works, let alone computer software. By contrast, the Ninth Circuit's decisions in *Sega*, 977 F.2d at 1524-27 (9th (extensive discussion of the second factor connecting fair use to *Baker v. Selden* and § 102(b)) and S*ony Comput. Entm't*, 203 F.3d at 602-05 (leading its discussion of fair use with the second fair use factor and affording it great significance), provide a far sounder footing for analyzing fair use in *Oracle v. Google*.

into play.[174] Google is seeking Supreme Court review of both the Federal Circuit's 2014 API copyrightability decision and its 2018 fair use decision.[175]

6. *Standards and Codes*

Copyright protection extends to any work of authorship fixed in a tangible medium of expression, subject to various limiting doctrines, such as the idea-expression dichotomy and fair use. Standard setting bodies generally promote access to their standards and codes. Sun (and later Oracle) published the Java API declarations. Their members typically wish to encourage widespread adoption of sponsored standards.

Some developers of standards seek to control access to their specifications. As reflected in the *Sega* case, Sega controlled the access codes for the Genesis game platform through trade secret law.[176] After Accolade successfully reversed engineered the interoperability code, Sega sought to bar its use by Accolade (and recover for copyright infringement). The Ninth Circuit held, however, that software code elements necessary for interoperability are unprotectable by copyright law.[177]

[174] *See* Menell, *supra* note 165, at 1518.

[175] *See* Google LLC v. Oracle America, Inc., U.S. Supreme Court No. 18-956, Petition for a Writ of Certiorari to the United States Court of Appeals for the Federal Circuit (Jan. 2019).

[176] *See* Sega Enters. v. Accolade, Inc., 977 F.2d 1510, 1532 (9th Cir. 1993).

[177] *See id.* at 1514 (referring to "unprotected functional elements of the program"); 1517 (referring to "functional requirements for Genesis compatibility"); 1522 (referring to "functional requirements for Genesis compatibility"); 1523 (noting that

> Accolade's identification of the functional requirements for Genesis compatibility has led to an increase in the number of independently designed video game programs offered for use with the Genesis console. It is precisely this growth in creative expression, based on the dissemination of other creative works and the unprotected ideas contained in those works, that the Copyright Act was intended to promote)

(citing Feist Publications, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991) (citing Harper & Row, 471 U.S. at 556–57)); 1524 (noting that "[i]n some circumstances, even the exact set of commands used by the programmer is deemed functional rather than creative for purposes of copyright. '[W]hen specific instructions, even though previously copyrighted, are the only and essential means of accomplishing a given task, their later use by another will not amount to infringement.' " (quoting CONTU REPORT, *supra* note 66, at 20)); 1525 (observing that

> [u]nder a test that breaks down a computer program into its component subroutines and sub-subroutines and then identifies the idea or core functional element of each, such as the test recently adopted by the Second Circuit in [*Computer Associates v. Altai*], many aspects of the program are not protected by copyright. In our view, in light of the essentially utilitarian

Various technical, building, and other standards development seek to control access to their work product principally to earn publication royalties. They contend that the royalty income provides vital funding for coordinating standard development, resulting in better formulated and maintained codes.[178]

Scholars have questioned the need for copyright protection to promote standards developments. Professor Paul Goldstein contends:

> [I]t is difficult to imagine an area of creative endeavor in which the copyright incentive is needed less. Trade organizations have powerful reasons stemming from industry standardization, quality control and self-regulation to produce these model codes; it is unlikely that, without copyright, they will cease producing them.[179]

The accessibility of edicts of law raises fundamental constitutional and policy questions.[180] Federal, state, and local laws, judicial opinions, and regulations incorporate these codes. The Copyright Act expressly exempts works of the federal government from copyright protection.[181] Court decisions on copyrightability of non-federal edicts of law have been mixed.

The Fifth Circuit held that model codes enter the public domain when they enter into law.[182] Building on that precedent, the Eleventh Circuit held that state law and the annotated compilation of such law are sufficiently law-like to

> nature of computer programs, the Second Circuit's approach is an appropriate one);

1526 (referring to "functional specifications" provided to clean room programmers); 1526 (observing that

> if disassembly of copyrighted object code is *per se* an unfair use, the owner of the copyright gains a *de facto* monopoly over the functional aspects of his work—aspects that were expressly denied copyright protection by Congress. 17 U.S.C. § 102(b). In order to enjoy a lawful monopoly over the idea or functional principle underlying a work, the creator of the work must satisfy the more stringent standards imposed by the patent laws)

(citing Bonito Boats, Inc. v. Thunder Craft Boats, Inc., 489 U.S. 141, 159–64 (1989)); 1527 (explaining that "[u]nder the Copyright Act, if a work is largely functional, it receives only weak protection . . . . Here, while the work may not be largely functional, it incorporates functional elements which do not merit protection").

[178] Letter from Jim Shannon, President, Nat'l Fire Protection Ass'n, to Michael White, Acting Dir., Office of the Fed. Register, The Nat'l Archives and Records Admin. (June 1, 2012) (on file with the Office of the Fed. Register, Request for Comments, Federal Register, Vol. 77, no. 38, NARA 12-0002 (February 27, 2012)).

[179] PAUL GOLDSTEIN, GOLDSTEIN ON COPYRIGHT § 2.5.2 (3d ed. 1996).

[180] *See Hearing on the Scope of Copyright Protection Before the H. Comm. on the Judiciary*, 113th Cong. 84–110 (2014) (statement of Carl Malamud, President, Public.Resource.Org).

[181] 17 U.S.C. § 105 (2018).

[182] *See* Veeck v. S. Bldg. Code Cong. Int'l, Inc., 293 F.3d 791 (5th Cir. 2002) (en banc).

be regarded as sovereign work constructively authored by the citizens and thus not copyrightable.[183]

By contrast, the First Circuit recognized that copyright law could potentially protect building codes.[184] The Ninth Circuit held that incorporation of a classification system (taxonomy) for medical procedures in Medicare and Medicaid regulations does not make them uncopyrightable.[185] Nonetheless, the court held that the copyright misuse doctrine limited the ability of the AMA to enforce its copyright against a health maintenance organization that used the taxonomy to comply with federal law.[186] Most recently, the D.C. Circuit overturned and remanded issuance of a permanent injunction barring a non-profit organization from distributing copies of technical standards produced by a private organization based on copyright and trademark grounds.[187] As the court noted, "[f]ederal, state, and local governments . . . have incorporated by reference thousands of these standards into law."[188] The court avoided a constitutional ruling by finding that the district court "failed to adequately consider whether, in certain circumstances, distributing copies of the law for purposes of facilitating public access could constitute transformative use."[189]

C. Trademark Protection, Unfair Competition Law, and False Advertising Protection

In contrast to patent, copyright, and trade secret protection—which seek to promote innovation—trademark, unfair competition law, and false advertising protection focus primarily on ensuring the integrity of the commercial marketplace.[190]

The federal Lanham Act as well as analogous state statutes and common law protects words, symbols, and other attributes, such as designs, slogans, and colors, that serve to identify the source of goods or services. Certification

[183] *See* Code Revision Comm'n v. Public.Resource.Org, Inc., 906 F.3d 1229, 1233, 1243–54 (11th Cir. 2018).

[184] *See* Bldg. Officials & Code Admin. v. Code Tech., Inc., 628 F.2d 730, 736 (1st Cir. 1980).

[185] Practice Mgmt. Info. Corp. v. Am. Med. Ass'n, 121 F.3d 516, 518–20 (9th Cir. 1997).

[186] *See infra* Section VI.A.1.

[187] *See* Am. Soc'y for Testing & Materials, et al. v. Public.Resource.Org, Inc., 896 F.3d 437 (D.C. Cir. 2018).

[188] *See id.* at 440.

[189] *See id.* at 450.

[190] *See generally* PETER S. MENELL, MARK A. LEMLEY & ROBERT P. MERGES, INTELLECTUAL PROPERTY IN THE NEW TECHNOLOGICAL AGE, VOL. II: COPYRIGHTS, TRADEMARKS, AND IP PROTECTIONS, ch. V (2018); Peter S. Menell & Suzanne Scotchmer, *Intellectual Property Law*, *in* 2 HANDBOOK OF LAW AND ECONOMICS ch. 19 (A. Mitchell Polinsky & Steven Shavell eds., 2007).

marks certify conformity with centralized standards. Collective marks connote that a product or service is manufactured or distributed by a member of a collective organization (e.g., Florists' Transworld Delivery Association (FTD)) or that a product or service provider is a member of a collective organization (e.g., American Automobile Association (AAA)). To receive trademark protection, a mark need not be new or previously unused, but it must represent a particular source of the good or service to consumers. It cannot merely describe the good (e.g., hotel) or represent a generic term (e.g., thermos) for the class of goods or services offered. Further, the identifying mark may not be a functional element of the product itself but must serve a purely identifying purpose. Trademarks do not expire, but continue in force unless their owner abandons them or they become generic.

Unlike patents or copyrights, trademarks do not directly protect the technology, good, or work, but rather prevent others from creating a likelihood of consumer confusion as to the source of goods. Thus, competitors may use the trademark of other companies in non-confusing ways, such as comparative advertising and descriptive usages. Furthermore, like copyright law, trademark law does not protect functional features of products.[191] Patent law provides the sole means of excluding competitors from utilitarian features of products. Similarly, trademark law cannot protect aesthetically functional features of goods or packaging. Thus, trademark law does not protect a red, heart-shaped box for packaging chocolates.[192] The Lanham Act and state laws prohibit false or misleading advertising.

In 1982, the Supreme Court applied the functionality doctrine in a case involving network effects.[193] Ives Laboratories manufactured and marketed a patented prescription drug using distinctively colored capsules: a blue capsule for its 200-mg dosage and a combination blue-red capsule for its 400-mg dosage.[194] Consumers and pharmacists came to associate the distinctive appearance of the capsules with the particular patented compound and dosages.[195] Thus, a consumer could identify whether they were taking the

[191] Kellogg Co. v. Nat'l Biscuit Co., 305 U.S. 111, 115, 122 (1938) (noting that the pillow-shaped form of shredded wheat biscuits reduces the cost of manufacturing the biscuits and affects their quality and therefore cannot serve as trade dress).

[192] RESTATEMENT (SECOND) OF TORTS § 742 cmt. a.

[193] *See* Inwood Labs., Inc. v. Ives Labs., Inc., 456 U.S. 844 (1982).

[194] *See id.* at 846–47.

[195] As noted in the *Inwood Labs* decision, most States enacted laws beginning in the early 1970's allowing pharmacists to substitute generic drugs for brand name drugs under certain conditions. *See id.* at 847 n.4 (citing Note, *Consumer Protection and Prescription Drugs: The Generic Drug Substitution Laws*, 67 KY. L.J. 384 (1978–1979)).

proper drug and dosage from its appearance. In that way, the packaging served as a simple language.

Following expiration of the utility patent, generic drug manufacturers marketed the chemical compound using the same color capsules.[196] Ives sued generic drug makers for indirect trademark infringement, alleging that they bore responsibility for pharmacists that mislabeled the source of the drugs. Many pharmacies distribute capsules in pharmacist-branded bottles.[197] The pharmacists violated trademark law by filling requests for Ives capsules with generic versions.[198] The generic companies only bore vicarious liability for the infringing acts of pharmacists, however, if they intentionally induced pharmacists to infringe the Ives trademark or if they continued to supply its product to pharmacists that it knew were engaging in infringement.[199]

In finding that Ives had not proven that the generic manufacturers were indirectly liable for trademark infringement, the Supreme Court observed that "a product feature is functional if it is essential to the use or purpose of the article or if it affects the cost or quality of the article."[200] A concurring opinion goes further, noting that

> a finding of functionality offers a complete affirmative defense to a contributory infringement claim predicated solely on the reproduction of a functional attribute of the product. A functional characteristic is 'an important ingredient in the commercial success of the product,' and, after expiration of a patent, it is no more the property of the originator than the product itself. It makes no more sense to base contributory infringement upon the copying of functional colors than on the petitioners' decision to use the same formulation of the drug, or even to market the generic substitute in the first place. To be sure, the very existence of generic drugs 'facilitates' illegal substitution. But Ives no longer has a patent for cyclandelate, 'and the defendants have a right to reproduce it as nearly as they can.' Reproduction of a functional attribute is legitimate competitive activity.[201]

Trademark and unfair competition regimes play a variety of roles in controlling and regulating information technology network markets by

[196] *See Inwood Labs.*, 456 U.S. at 847.

[197] *See id.*

[198] *See id.* at 854–55 (recognizing that "pharmacists who mislabeled generic drugs with Ives' registered trademark violated [Lanham Act] § 32").

[199] *See id.*

[200] *Id.* at 850 n.10.

[201] *Id.* at 862–63 (White, J., concurring) (citations omitted).

enabling platform sponsors to regulate the usage of terms and symbols that signal interoperability and compatibility with particular standards and interfaces.[202]

Platform sponsors and standard setting organizations routinely establish certification and collective markets and use trademark law to police use of these designations. As noted above, Sun Microsystems (and now Oracle Corporation) uses the Java TCK test as well as certification marks to ensure that products using the Java trademark meet WORA interoperability standards. In the mid to late 1990s, Sun used the "100% Pure Java" initiative to establish Java as a de facto industry standard.[203] Sun successfully sued Microsoft for violating its agreement not to adhere to Java's standardized application environment and compliance tests so as to ensure interoperability.[204]

Platform sponsors have used trademark and false advertising law to combat confusing product names or packaging and police compatibility and interoperability claims. Apple Computer, for example, successfully prevented a competitor from using the term "Pineapple" for its clone device.[205] As another example, Hewlett-Packard blocked an ink refiller from using confusingly similar packaging for replacement cartridges.[206]

In an interesting application of trademark's genericide doctrine, Intel Corporation sought to protect the "x86" suffix from confusing use by a competitor. The court determined, however, that the "x86" designation had become generic among buyers and sellers of microprocessor chips.[207] Consequently, Intel designated its fifth generation design the Pentium. By contrast, notwithstanding the serious questions a court raised about whether "Windows" was generic for a graphical user interface,[208] Microsoft obtained federal registration for the Windows term. Google has successfully fended off claims that "google" has become a generic term for Internet search.[209]

[202] *See* MARK A. LEMLEY, PETER S. MENELL, ROBERT P. MERGES, PAMELA SAMUELSON & BRIAN W. CARVER, 1 SOFTWARE AND INTERNET LAW ch. 4 (4th ed. 2011).

[203] Paul Floren, *Sun's Java: Can It Burn Microsoft?*, N.Y. TIMES (Jan. 20, 1997), http://www.nytimes.com/1997/01/20/business/worldbusiness/20iht-java.t.html [https://perma.cc/Y3UQ-N2TW].

[204] *See* Sun Microsystems, Inc. v. Microsoft Corp., 87 F. Supp. 2d 992 (N.D. Cal. 2000).

[205] *See* Apple Comput., Inc. v. Formula Int'l, Inc., 562 F. Supp. 775 (C.D. Cal. 1983).

[206] *See* Hewlett-Packard Co. v. Nu-Kote Int'l, Inc., No. Civ.A.C94–20647JW (EA, 2000 WL 33992123 (N.D. Cal. 2000).

[207] *See* Intel Corp. v. Advanced Micro Devices, Inc., 756 F. Supp. 1292 (N.D. Cal. 1991).

[208] *See* Microsoft Corp. v. Lindows.com, Inc., No. C01-2115C, 2002 WL 32085606 (W.D. Wash. 2002).

[209] *See* Elliott v. Google, Inc., 860 F.3d 1151 (9th Cir. 2017).

Platform sponsors and complementary product manufacturers have used trademark and false advertising law to police use of compatibility and interoperability claims. In *Princeton Graphics Operating, L.P. v. NEC Home Electronics (U.S.A.), Inc.*,[210] the court applied a restrictive definition of compatibility because of the importance of precise definitions in the computer industry.[211]

In another interesting application of trademark law's functionality doctrine, the Ninth Circuit declined to allow Sega to use trademark law to prevent Accolade from selling interoperable products that displayed Sega's trademark as part of its lock-out code.[212] The basis for the trademark claim was that the initialization code prompted a visual display for approximately three seconds that read "PRODUCED BY OR UNDER LICENSE FROM SEGA ENTERPRISES LTD."[213] The court rejected the false labeling claim as inconsistent with the purposes of the Lanham Act.[214] It also held that Sega could not use trademark law to prevent competitors from marketing interoperable devices if the software design required display of what might otherwise be confusing trademark information.[215] The court ruled that Sega failed to prove the existence of a feasible alternative to using the lock-out code that produced the misleading label.[216] Furthermore, Accolade had placed text on its packaging materials disclaiming any association with Sega.[217]

D. Patent Protection

Patents have long provided the potential for exclusive rights for network technologies. For example, Alexander Graham Bell, who edged out Elisha Gray in a patent race over the telephone, gained monopoly control over the quintessential network technology.[218] As the Supreme Court noted in *Dolbear v. Am. Bell Tel. Co.*,[219] although an inventor's claim might practically preempt

[210] 732 F. Supp. 1258 (S.D.N.Y. 1990).

[211] *See* Creative Labs, Inc. v. Cyrix Corp., 42 U.S.P.Q.2d 1872 (N.D. Cal. 1997) (following the *Princeton Graphics Operating* restrictive definition of compatibility and finding that a product advertisement asserting compatibility with a competing product must support the same functions).

[212] *See* Sega Enters., Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992).

[213] *See id.* at 1515.

[214] *See id.* at 1528–30.

[215] *See id.* at 1530–32.

[216] *See id.*

[217] *See id.* at 1529, 1532 n.11.

[218] *See generally* Robert V. Bruce, Bell: Alexander Graham Bell and the Conquest of Solitude (1990).

[219] 126 U.S. 1, 535 (1888).

all use of a discovery for the duration of the patent, this fact will "show more clearly the great importance of his discovery, [] it will not invalidate [the preempting] patent." Patents tracing back to Guglielmo Marconi wireless communications technology played a central role in the development of the radio and television industries.[220] Xerox controlled the photocopying industry for several decades in the mid-20th century. Intel built its microprocessor juggernaut on patents. Other network technology industries—from modems[221] to cell phones (Code Division Multiple Access (CDMA))[222]—were built on patent portfolios. Concern over patents affects many standard-setting processes.[223]

The extent to which patents enable control of network technologies depends on a range of factors, including the extent to which the patent controls network features (patent scope), the effective duration of patent protection, licensing structures (including patent pools), [224] and antitrust constraints.

The advent of computer software introduced several additional complicating factors. As courts limited copyright protection for network features of computer software and the Federal Circuit expanded patent eligibility for software-related inventions in the 1990s, the patent system emerged as a battleground for software-related network technologies. Patent law's higher protection threshold compared to other intellectual property modes seeks to ensure that trivial advances remain available to the public while potentially providing substantial advances robust protection, thereby motivating platform developers to take on the challenge of overcoming the excess inertia of entrenched, but obsolete, platforms. Patent law's disclosure requirements enable the public to learn from technological advances. Nonetheless, patent protection's twenty-year duration, although far shorter than copyright protection, might still be excessive for software technologies.[225] The uncertain scope of patent protection also poses some concern. Patent remedies can be especially strong, although standard-setting processes have

[220] *See generally* HUGH G.J. AITKEN, THE CONTINUOUS WAVE: TECHNOLOGY AND AMERICAN RADIO 1900–32 (1985).

[221] *See* Neil Gandal, Nataly Gantman & David Genesove, *Intellectual Property and Standardization Committee Participation in the US Modem Industry*, *in* 1 STANDARDS AND PUBLIC POLICY 208 (Shane Greenstein & Victor Stango eds., 2007).

[222] *See generally* DAVE MOCK, THE QUALCOMM EQUATION: HOW A FLEDGLING TELECOM COMPANY FORGED A NEW PATH TO BIG PROFITS AND MARKET DOMINANCE (2005).

[223] *See generally* Contreras, *supra* note 26.

[224] *See* Michael Mattioli, *Empirical Studies of Patent Pools*, *in* 2 RESEARCH HANDBOOK ON THE ECONOMICS OF INTELLECTUAL PROPERTY (Peter S. Menell & David L. Schwartz eds., 2019); *see also*, Mattioli, *supra* note 13.

[225] *See* Menell, *supra* note 23, at 1364–65.

tempered their effects and promoted collaboration. Finally, design patent protection has recently added a new weapon to the network technology arsenal.

This Section examines patent protection for network technologies. It emphasizes the most salient and contested area: computer software. Section 1 traces the evolution of patent protection for software-related inventions. Section 2 examines the complicated scope of patent protection. Section 3 discusses patent licensing. Section 4 explores patent remedies. Section 5 examines design patents and their emergence in network markets.

1. *Patentability Requirements*

The Patent Act sets forth five patentability requirements: (1) patentable subject matter; (2) utility; (3) novelty; (4) nonobviousness; and (5) disclosure.[226] Two of these requirements have been particularly pertinent to network industries: subject matter eligibility and nonobviousness.

a) Subject Matter Eligibility

As noted above, the patent system has long afforded protection for network and systems technologies, ranging from the telephone to wireless communication and xerography. These technologies fit comfortably within the traditional scope of patent protection. The patent system has, however, struggled to accommodate software-related inventions. As illustrated above, APIs and other software technologies are increasingly important in network industries.

Notwithstanding that the patent statute expressly authorizes patenting of processes and machines,[227] the availability of patent protection for software-related inventions has been in flux since the beginning of the computer age. The issue emerged in the 1960s as computer systems became more versatile, software languages developed, and computer programming emerged from the shadow of electrical engineering. The Patent Office struggled to fit software inventions within the traditional classification system and struggled to keep up with the tremendous volume of prior art being generated. In 1965, President Johnson appointed a commission to assess the overall efficacy of the patent system.[228] In recommending that Congress exclude computer programs from patent eligibility, the Commission of government officials, leading scientists, and representatives of industry (including IBM), noted that "the creation of

[226] 35 U.S.C. §§ 101, 102, 103, 112 (2018); *see* MENELL ET AL., *supra* note 2, ch. 3.

[227] *See* 35 U.S.C. § 101.

[228] Executive Order No. 11,215, 30 Fed. Reg. 4661 (1965).

programs has undergone substantial and satisfactory growth in the absence of patent protection" and that "copyright protection for programs is presently available."[229] But as discussed above, copyright excluded protection for functional features of expressive works.

Congress did not act on this recommendation, and the eligibility of software-related inventions fell to the Patent Office and the courts. Although granting a smattering of software-related inventions in the mid to late 1960s, the Patent Office took a skeptical view of software eligibility. This in part reflected concerns that about the PTO's ability to examine this new and rapidly developing technological field.

The Supreme Court was soon brought into the fray. An inventor challenged the PTO's rejection of his claim to an algorithm that converted binary-coded decimal numerals into pure binary numerals on subject matter grounds.[230] The Court held that "[p]henomena of nature, though just discovered, mental processes, and abstract intellectual concepts are not patentable, as they are the basic tools of scientific and technological work."[231] The court noted, however, it was not categorically excluding software-related inventions from patent eligibility.[232] Yet six years later, the Court ruled that even newly discovered algorithms should be treated as in the prior art, rendering software claims ineligible unless they contained some other inventive concept.[233] The Supreme Court reversed course in 1981, holding that software claims should be viewed as a whole and that the touchstone for patentability of a process embodying a mathematical formula was whether there was significant post-solution activity that is "transforming or reducing an article to a different state or thing."[234]

Over the ensuing twenty-five years, the Court of Appeals for the Federal Circuit loosened patent eligibility limitations. Building on *Diehr*, the Federal Circuit chipped away at the post-solution activity necessary to bring software-related claims within § 101.[235] In 1998, the Federal Circuit held that business methods were eligible for patent protection so long as they produced a "useful, concrete and tangible result."[236]

[229] U.S. COMM. ON THE JUDICIARY, REPORT OF THE PRESIDENT'S COMMISSION ON THE PATENT SYSTEM, S. REP. DOC. NO. 90-5 (1967).

[230] *See* Gottschalk v. Benson, 409 U.S. 63 (1972).

[231] *Id.* at 67.

[232] *Id.*

[233] *See* Parker v. Flook, 437 U.S. 584, 594 (1978).

[234] Diamond v. Diehr, 450 U.S. 175, 183, 188–89, 191–92 (1981).

[235] *See In re* Alappat, 33 F.3d 1526 (Fed. Cir. 1994) (holding that the display of data on a computer screen could suffice).

[236] *See* State St. Bank & Trust Co. v. Signature Fin. Grp., 149 F.3d 1368, 1373 (Fed. Cir. 1998) (quoting *In re* Alappat).

In the aftermath of the Federal Circuit's *State Street Bank* decision, the PTO shifted its position from skepticism about expansive patent eligibility to openness and even enthusiasm. Patents for software and business methods flooded the PTO. Entrepreneurs and venture capitalists saw patenting as a valuable tool for developing (or at least claiming) Internet businesses. The late 1990s witnessed unprecedented growth of start-up businesses based on speculative initial public offerings secured, in part, on patent portfolios.

The bursting of the Internet (dot-com) stock bubble in 2000 produced a dramatic shakeout. Bankruptcies and, subsequently, the auctioning and trading of Internet-related patents, became widespread. Entities whose sole purpose was to assert these patents emerged. Patent holding companies and non-practicing entities sought to monetize their Internet patents, often purchased at bankruptcy auctions. Lawsuits by patent assertion entities produced a tidal wave of patent validity challenges as well as calls by Silicon Valley companies, policymakers, and scholars for policy reform.

These concerns led the Federal Circuit to reinvigorate patent eligibility limitations.[237] In an en banc ruling, the Federal Circuit synthesized the Supreme Court's *Benson*, *Flook*, and *Diehr* precedents into the "machine-or-transformation test": a claimed process is patent-eligible under § 101 if it is tied to a particular machine or if it transforms a particular article into a different state or thing.[238] Applying this test, the Federal Circuit affirmed the Patent Office's rejection of a claim for a method for managing the consumption risk costs of a commodity. The Supreme Court upheld the Federal Circuit's decision, although it characterized the machine-or-transformation test as a "useful and important clue, an investigative tool, for determining whether some claimed inventions are processes under § 101," but too rigid a test of the Patent Act's broad statutory definition of "process."[239] The Court declined to rule that business methods are categorically ineligible for patent protection.[240]

Two years later, the Supreme Court revived the *Flook* decision's rule that for a claim embodying a natural discovery or algorithm to be eligible for patentability, it must contain a sufficiently inventive concept beyond the

[237] *See, e.g.*, *In re* Nuijten, 500 F.3d 1346 (Fed. Cir. 2007) (holding that a watermarked electromagnetic signal does not fall into any of the four categories of patent-eligible subject matter); *In re* Comiskey, 554 F.3d 967 (Fed. Cir. 2009) (affirming rejection of a business method patent under § 101 as merely relying on mental steps).

[238] *In re* Bilski, 545 F.3d 943, 961 (Fed. Cir. 2008) (en banc).

[239] Bilski v. Kappos, 561 U.S. 593, 604 (2010); 35 U.S.C. § 100(b) (2018).

[240] *See* Peter S. Menell, *Forty Years of Wandering in the Wilderness and No Closer to the Promised Land:* Bilski*'s Superficial Textualism and the Missed Opportunity to Return Patent Law to Its Technology Mooring*, 63 STAN. L. REV. 1289, 1299–304 (2011).

natural law or algorithm, even where the patentee discovered the natural law or algorithm.[241] These decisions have dramatically shifted the patent-eligibility landscape, resulting in the invalidation of a vast swath of software-related claims and eliminating patent protection for pure business methods. The decisions have also reduced the availability of patent protection for software-based network technologies.

b) Nonobviousness

To ensure that patents are not granted to routine or conventional applications of known principles, the Patent Act stipulates that a patent for an invention may not be obtained if "the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the [invention was made] to a person having ordinary skill in the art to which the claimed invention pertains."[242] This requirement has long been difficult to apply due to the difficulty of ignoring the fact of the claimed invention. To avoid such hindsight bias, the Federal Circuit interpreted § 103 to require that the prior art teach, suggest, or motivate ordinary skilled artisans to combine prior art references to achieve the claimed invention. Absent such evidence, the claimed invention was nonobvious.[243] While such suggestions can be relatively common in scientific publications—through cross-references of other publications—they are not readily found in more commercial and applied fields, such as software engineering. Software products do not typically cross-reference other products. As a result, many seemingly obvious inventions from the standpoint of common knowledge were able to clear the Federal Circuit's nonobviousness test.

As software patent litigation exploded following the burst of the Internet bubble in 2000, the Federal Circuit's standard for determining whether an invention was sufficiently inventive came under scrutiny. In *KSR International Co. v. Teleflex Inc.*,[244] the Supreme Court tightened the nonobviousness standard by holding that the teaching-suggestion-motivation test was too rigid:

> When there is a design need or market pressure to solve a problem and there are a finite number of identified, predictable solutions, a person of ordinary skill has good reason to pursue the known options within his or her technical grasp. If this leads to the

[241] *See* Mayo Collaborative Servs. v. Prometheus Labs., Inc., 132 S. Ct. 1289 (2012); Alice Corp. v. CLS Bank International, 134 S. Ct. 2347 (2014).

[242] 35 U.S.C. § 103.

[243] *See, e.g.*, Teleflex, Inc. v. KSR Int'l Co., 119 Fed. App'x 282, 285 (Fed. Cir. 2005), *rev'd*, 550 U.S. 398 (2007); *In re* Dembiczak, 175 F.3d 994, 998 (Fed. Cir. 1999); *In re* Bergel, 292 F.2d 955, 956–57 (C.C.P.A. 1961) (predecessor court to the Federal Circuit).

[244] 550 U.S. 398 (2007).

> anticipated success, it is likely the product not of innovation but of ordinary skill and common sense. In that instance the fact that a combination was obvious to try might show that it was obvious under § 103.[245]

The *KSR* decision raised the patentability bar, especially for software-related technologies for which market factors and advances in collateral technologies are likely to drive new products and processes.

2. *Scope*

The extent to which patents control network technologies depends upon the scope of the patent claims. Pioneering patents can stake broad claims without fear of being anticipated by prior art, whereas incremental inventions in crowded technology fields only garner narrow protection. Moreover, pioneering inventors can often develop improvement patents that expand their control and duration of protection. Xerox successfully followed this strategy to monopolize the photocopying industry for several decades.[246] The resulting "patent thicket" delayed entry into the plain paper copy industry.

Software patentees have used broad, vague functional claim language to obtain broad coverage for their inventions.[247] By avoiding the statutory phrases "means" or "step" in their claims—which limit the scope of their claims to the particular embodiments in the specification and "equivalents thereof"[248]—and instead using broad terms that lack structural limits such as "module," patent drafters have sought to control all software solutions to particular technological problems.[249] Such claims have caused substantial problems in the Internet Age, and have resulted in a proliferation of demand letters, costly litigation, and nuisance value settlements.

The courts and the PTO have sought to rein in these problems. The Supreme Court invigorated the claim indefiniteness doctrine, enforcing the patent statute's requirement to "particularly point[] out and distinctly claim[]

[245] *Id.* at 421.

[246] *See* F. M. Scherer, *Antitrust, Efficiency, and Progress*, 62 N.Y.U. L. REV. 998, 1016–17 (1987); *see generally* Timothy F. Bresnahan, *Post-Entry Competition in the Plain Paper Copier Market*, 75 AM. ECON. REV. 15 (1985).

[247] *See* Peter S. Menell & Michael J. Meurer, *Notice Failure and Notice Externalities*, 5 J. LEGAL ANALYSIS 1, 33 (2013); FTC REPORT, *supra* note 38.

[248] 35 U.S.C. § 112(f).

[249] *See generally* Mark A. Lemley, *Software Patents and the Return of Functional Claiming*, 2013 WIS. L. REV. 905 (2013).

the subject matter" sought to be patented.[250] The Federal Circuit has interpreted claim terms like "module" and other vague terms (which it refers to as "nonce" words) to invoke the limitations of § 112(f).[251] This interpretation limits claim scope to the embodiments in the specification and equivalents thereof. Further upstream, the Patent Office is pursuing administrative efforts to improve claim clarity.[252]

3. *Licensing*

Patent licensing plays a critical role in many network industries. Patents afford patent owners the power to prevent others from making, using, offering to sell, selling, or importing the patented invention in the United States during the term of the patent.[253] They do not, however, ensure that patentees can practice their own patented invention. The owner of a patent that improves on patented technologies controlled by others would need a license from the upstream patent owner to make, use, or sell the improvement. Licensing provides the key.

Many network technologies employ patented technologies. Several distinctive licensing issues have developed to address network effects: (a) standard setting and commitments to license patents on FRAND terms; (b) insurance pools and license on transfer commitments; and (c) GPL viral license commitments. Overreaching licensing provisions can raise misuse and antitrust issues addressed in Part VI.

a) Standard-Setting and FRAND Commitments

SSOs seek to lessen the tension between employing the best technological solutions in industry standards and ensuring widespread access to standards by requiring members to disclose standard-essential patents (SEPs) and license them on FRAND terms.[254] Most SSOs, however, have not expressly barred injunctive relief or set FRAND licensing schedules. In 2015, the Institute of Electrical and Electronics Engineers (IEEE) barred its members holding

[250] Nautilus, Inc. v. Biosig Instruments, Inc., 134 S. Ct. 2120, 2124 (2014); 35 U.S.C. § 112(b); Menell & Meurer, *supra* note 247, at 33.

[251] *See* Williamson v. Citrix Online LLC, 792 F.3d 1339, 1350 (Fed. Cir. 2015) (en banc).

[252] United States General Accountability Office, *Intellectual Property: Patent Office Should Define Quality, Reassess Incentives, and Improve Clarity*, GAO-16-490 (July 20, 2016), https://www.gao.gov/assets/680/678113.pdf [https://perma.cc/LXV3-MDDG]; Peter S. Menell, *It's Time to Make Vague Software Patents More Clear*, WIRED (Feb. 7, 2013), http://www.wired.com/opinion/2013/02/its-time-to-make-vague-software-patents-more-clear/ [https://perma.cc/94KB-ENTY].

[253] *See* 35 U.S.C. § 271.

[254] *See* Contreras, *supra* note 26, at 23; *see also* NAT'L RESEARCH COUNCIL OF THE NAT'L ACADS., *supra* note 36; Mattioli, *supra* note 13.

patents covering IEEE standards from seeking or threatening to seek injunctions or exclusion orders against potential licensees who are willing to negotiate licenses.[255]

b) Insurance Pools and License on Transfer (LOT) Commitments

In response to widespread assertion of patents by non-practicing entities following the bursting of the Internet bubble in early 2000, several enterprises emerged to reduce patent risk.[256] Since 2008, RPX (Rational Patent Exchange) Corporation has functioned as a consortium of technology companies that acquires patents that pose potential risks. RPX has promised not to assert patents in its portfolio.[257]

As a further pre-commitment strategy to prevent patent holdup, a growing number of technology companies have promised not to assert their patents under specified conditions.[258] Google has led an initiative whereby companies agree to prevent their patents from ever being used by a non-practicing entity (NPE) against other member companies through a license on transfer (LOT) pledge.[259] The LOT network produces a network benefit. As more companies join the pact, the freedom to be insulated from NPE patent assertion entities expands.

c) GPL 3.0

As noted earlier, patents did not play a substantial role in the software industry until the mid-1990s, after the GPL (1989) and the GPL 2 (1991) were established. Although neither version of the GPL expressly licensed patents, the Free Software Foundation took the position that the GPL 2 created an implied license.[260]

GPL 3.0 took aim at this issue. Section 10 provides that the licensee

[255] *IEEE-SA Standards Board Bylaws*, INST. OF ELECTRICAL AND ELECTRONICS ENGINEERS § 6 (2015), http://standards.ieee.org/develop/policies/bylaws/approved-changes.pdf [https://perma.cc/SEF8-PL2Y] (last visited Oct. 21, 2018).

[256] *See* James M. Rice, *The Defensive Patent Playbook*, 30 BERKELEY TECH. L.J. 725, 752–53 (2015).

[257] *See RPX Corporation*, WIKIPEDIA, https://en.wikipedia.org/wiki/RPX_Corporation [https://perma.cc/K6JK-EUVE].

[258] *See* Rice, *supra* note 256, at 747–53.

[259] *Id.* at 768–69; *cf.* Jason Schultz & Jennifer M. Urban, *Protecting Open Innovation: The Defensive Patent License as a New Approach to Patent Threats, Transaction Costs, and Tactical Disarmament*, 26 HARV. J.L. & TECH. 1, 64–65 (2012) (proposing a precursor to the license on transfer model).

[260] MEEKER, *supra* note 12, at 127.

> may not impose any further restrictions on the exercise of the rights granted or affirmed under this License. For example, you may not impose a license fee, royalty, or other charge for exercise of rights granted under this License, and [the licensee] may not initiate litigation (including a cross-claim or counterclaim in a lawsuit) alleging that any patent claim is infringed by making, using, selling, offering for sale, or importing the Program or any portion of it.[261]

Section 11 goes further: each "contributor" to code governed by GPL 3.0 grants

> a non-exclusive, worldwide, royalty-free patent license under the contributor's essential patent claims, to make, use, sell, offer for sale, import and otherwise run, modify and propagate the contents of its contributor version.[262]

That provision defines a contributor's "essential patent claims" to include

> all patent claims owned or controlled by the contributor, whether already acquired or hereafter acquired, that would be infringed by some manner, permitted by this License, of making, using, or selling its contributor version.[263]

Section 11 does not extend to "claims that would be infringed only as a consequence of further modification of the contributor version."[264]

Section 11 further provides that a licensee who is aware of a patent license governing by GPL 3.0 code must make the corresponding source code to run the object code and modify the work publicly available or extend the patent license to downstream recipients.[265] Alternatively, the licensee must deprive itself of the benefit of the license. Section 11 further includes a non-discrimination provision ensuring that any patent licenses are extended to all recipients of the GPL 3.0 work and works based on it.[266]

These provisions pose several serious concerns to many commercial software developers.[267] For example, many patent litigation settlements provide only limited, non-sublicenseable, and possibly royalty-bearing rights that would not comply with GPL 3.0 requirements. Thus, commercial enterprises have been reluctant to embrace GPL 3.0. As of February 2017,

[261] *GNU General Public License*, *supra* note 151, at § 10.

[262] *Id.* at § 11.

[263] *Id.*

[264] *Id.*

[265] *See id.*

[266] *See id.*

[267] MEEKER, *supra* note 12, at 129–30.

GPL 3.0 was the fourth most widely adopted open source license (8% of open source projects), behind the MIT License (a simple permissive) (31%), GPL 2.0 (18%), and Apache 2.0 (15%).[268]

4. *Remedies*

Patent remedies play a critical role in the control of network technologies that are subject to patent assertions. The proliferation of software patents and litigation in the Internet Age generated tremendous exposure for network industry companies, leading to calls for statutory reform of patent remedies.[269]

a) Injunctive Relief

The patent right—the right to exclude others from practicing the patented technology—has historically been protected by injunctive relief. Courts traditionally viewed patent rights like other property interests and routinely protected them through a "property" rule—barring transgressors from trespassing or using the "property."[270] Thus, for most of the history of patent law, courts awarded a permanent injunction as the prospective infringement remedy absent extraordinary circumstances.[271]

The embrace of software and business method patents during the dot-com bubble of the mid- to late 1990s gave way to concerns about injunctions threatening major technology companies in the aftermath of the NASDAQ

[268] *Top Open Source Licenses*, BLACK DUCK BY SYNOPSYS, https://www.blackducksoftware.com/top-open-source-licenses [https://perma.cc/3TR7-3T42] (last visited Apr. 7, 2019); Ayala Goldstein, *Top 10 Open Source Licenses in 2018: Trends and Predictions*, WHITE SOURCE (Dec. 3, 2018) https://resources.whitesourcesoftware.com/blog-whitesource/top-open-source-licenses-trends-and-predictions [https://perma.cc/29WN-FVYH] (noting that use of permissive open source licenses are on the rise; reporting that in 2018, 64% of open source components have permissive licenses, an 8% rise over 2017).

[269] *See* FTC REPORT, *supra* note 38; *see also* FED. TRADE COMM'N, TO PROMOTE INNOVATION: THE PROPER BALANCE OF COMPETITION AND PATENT LAW AND POLICY (2003).

[270] *See* MercExchange, LLC v. eBay, Inc., 401 F.3d 1323, 1338 (Fed. Cir. 2005) ("Because the 'right to exclude recognized in a patent is but the essence of the concept of property,' the general rule is that a permanent injunction will issue once infringement and validity have been adjudged.") (citing Richardson v. Suzuki Motor Co., 868 F.2d 1226, 1246–47 (Fed. Cir. 1989)), *vacated and remanded*, eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388 (2006).

[271] *See id.* (noting that "courts have in rare instances exercised their discretion to deny injunctive relief in order to protect the public interest") (citing Rite–Hite Corp. v. Kelley, Inc., 56 F.3d 1538, 1547 (Fed. Cir. 1995)); *see also* Roche Prods., Inc. v. Bolar Pharm. Co., 733 F.2d 858, 865–66 (Fed. Cir. 1984) ("[S]tandards of the public interest, not the requirements of private litigation, measure the propriety and need for injunctive relief."); Milwaukee v. Activated Sludge, Inc., 69 F.2d 577 (7th Cir.1934) (declining to issue an injunction where the shutdown of a sewage disposal plant posed public health danger).

market crash in the early 2000s. Patents that had been acquired to attract venture capital were auctioned off in bankruptcy sales to patent monetization entities.[272] The proliferation of demand letters and patent lawsuits led scholars, technology companies, policymakers, and jurists to reconsider the traditional view of patents as property interests that deserve near-automatic injunctive relief.[273] The costs of identifying patent holders, negotiating among potentially hundreds of patent holders, and the disruption and delay of litigation created leverage for patent owners. The threat of injunctive relief and high monetary damages enabled holders of dubious patents to extract unwarranted and disproportionate value.

In a watershed decision, the Supreme Court ruled in *eBay, Inc. v. MercExchange, LLC*[274] that the award of injunctive relief in patent cases turns on equitable balancing of the traditional equitable factors associated with preliminary relief: (1) whether the harm is irreparable, (2) adequacy of monetary damages to compensate for the harm, (3) balance of hardships between the parties, and (4) the public interest.[275] The *eBay* decision has changed patent remedies dramatically. Seaman finds that the overall rate of permanent injunctions being ordered as a remedy for patent infringement has dropped from near 100% to 72.5%.[276] The drop is most significant in software cases (53%). Patent assertion entities obtained permanent injunctions in just 16% of their victories.

Courts take SSO FRAND commitments into account in evaluating requests for injunctive relief under the *eBay* standard. Although many SSO policies do not expressly address whether SEP owners can seek injunctive

[272] *See* Colleen V. Chien, *From Arms Race to Marketplace: The Complex Patent Ecosystem and Its Implications for the Patent System*, 62 HASTINGS L.J. 297, 304–06 (2010); KEVIN G. RIVETTE & DAVID KLINE, REMBRANDTS IN THE ATTIC: UNLOCKING THE HIDDEN VALUE OF PATENTS (2000).

[273] *See* William F. Lee & A. Douglas Melamed, *Breaking the Vicious Cycle of Patent Damages*, 101 CORNELL L. REV. 399, 435–36 (2016); Peter S. Menell, *The Property Rights Movement's Embrace of Intellectual Property: True Love or Doomed Relationship?*, 34 ECOLOGY L.Q. 713, 718 (2007); Peter S. Menell, *Governance of Intellectual Resources and Disintegration of Intellectual Property in the Digital Age*, 26 BERKELEY TECH. L.J. 1523, 1549–50 (2011); *see generally* Lemley & Shapiro, *supra* note 30.

[274] See Mark P. Gergen, John M. Golden, & Henry E. Smith, *The Supreme Court's Accidental Revolution? The Test for Permanent Injunctions*, 112 COLUM. L. REV. 203, 208–09 (2012).

[275] 547 U.S. 388 (2006).

[276] *See* Christopher B. Seaman, *Permanent Injunctions in Patent Litigation After* eBay*: An Empirical Study*, 101 IOWA L. REV. 1949, 1982–83 (2016).

relief or exclusion orders, courts consider FRAND commitments in weighing the irreparable harm prong of the *eBay* equitable relief test.[277]

The *eBay* decision does not, however, leave the patent owner without a prospective remedy. The court will fashion a prospective monetary damage measure, such as a running royalty or a permanent damage amount—essentially a compulsory license.[278] The *eBay* decision has led to a rise in patent enforcement filings at the International Trade Commission, which enforces infringement findings with exclusion orders barring importation of infringing articles.[279]

b) Monetary Relief

The Patent Act authorizes the award of "damages adequate to compensate for the infringement, but in no event less than a reasonable royalty for the use made of the invention by the infringer."[280] Thus, patentees can recover lost profits or a reasonable royalty resulting from infringing activity. The Patent Act further authorizes judges to increase damages awards up to three times the compensatory level where the infringer has acted willfully or recklessly.[281] Policymakers and scholars see the goal of patent damages to restore the parties to the position they would have achieved had they negotiated a patent license before the infringement occurred.[282]

Patent law has long struggled to deal with apportioning patent value when a patent covers only one component of a larger product or system.[283] The problem has become particularly acute in platform technologies involving multiple components and patented technologies. The serial nature of patent

[277] Apple Inc. v. Motorola, Inc., 757 F.3d 1286, 1331–32 (Fed. Cir. 2014) (noting that absent unusual circumstances, such as an infringer refusing a FRAND royalty or unreasonably delaying negotiations, it will be difficult for a patent owner subject to a FRAND commitment to establish irreparable harm or that damages are not an adequate remedy, and that even when an infringer has refused to accept any license offer, that does not necessarily justify injunctive relief); *cf.* Apple Inc. v. Samsung Elecs. Co., 809 F.3d 633 (Fed. Cir. 2015) (emphasizing right to exclude and the importance of injunctions).

[278] *See* Paice LLC v. Toyota Motor Corp., 504 F.3d 1293, 1314 (Fed. Cir. 2007); Christopher B. Seaman, *Ongoing Royalties in Patent Cases After eBay: An Empirical Assessment and Proposed Framework*, 23 TEX. INTELL. PROP. L.J. 203 (2015).

[279] *See* Colleen V. Chien & Mark A. Lemley, *Patent Holdup, the ITC, and the Public Interest*, 98 CORNELL L. REV. 1, 2–3 (2012).

[280] 35 U.S.C. § 284 (2018).

[281] *See id.*; Halo Elecs., Inc. v. Pulse Elecs., Inc., 136 S. Ct. 1923 (2016).

[282] *See* Lee & Melamed, *supra* note 77, at 392.

[283] *See* Cincinnati Car Co. v. New York Rapid Transit Corp., 66 F.2d 592, 593 (2d Cir. 1933) (Learned Hand, J.) (observing that the allocation of profits among multiple components "is in its nature unanswerable").

litigation, the economic complexity of multi-component products, and court-imposed time limits on the presentation of evidence make it difficult for juries to apportion value among multiple components and factors driving market demand for infringing products.[284]

In theory, a wide range of royalty bases can be used with appropriately calibrated royalty rates to account for the myriad factors affecting consumer demand. In practice, however, the open-ended nature of the inquiry,[285] can lead to a very large royalty range across comparable cases. The Federal Circuit has sought to rationalize awards by using the smallest saleable patent-practicing unit (SSPPU), as opposed to the entire market value of the product or system, as the royalty base.[286]

As noted above, SSOs have sought to alleviate the tension between technological progress and widespread access to standards by requiring members to disclose SEPs during the standard-setting process and license them to standards implementers on FRAND terms.[287] Nonetheless, the valuation of SEPs is difficult, especially when industry standards encompass multiple technologies and hundreds of patents. The challenge lies in separating the value of the particular technologies and patents from the often tremendous value from standardization, which is attributable to network effects. Once consumers adopt a product, they become locked-in to the standard to varying degrees. This can provide patentees with tremendous leverage. Courts have surmounted this challenge by interpreting the principal goal of standard-setting agreements to be widespread adoption of the standard and barring FRAND licensors from capturing the coordination and network value of the standard.[288]

[284] *See* Stuart Graham, Peter S. Menell, Carl Shapiro, & Tim Simcoe, *Final Report of the Berkeley Center for Law & Technology Patent Damages Workshop*, 25 TEX. INTELL. PROP. L.J. 115, 127–28 (2017).

[285] *See* Georgia-Pacific Corp. v. U.S. Plywood Corp., 318 F. Supp. 1116 (S.D.N.Y. 1970) (identifying fifteen factors).

[286] *See* LaserDynamics Inc. v. Quanta Comput., Inc., 694 F.3d 51 (Fed. Cir. 2012); Cornell Univ. v. Hewlett-Packard Co., 609 F. Supp. 2d 279 (N.D.N.Y. 2009) (Rader, J., sitting by designation); Uniloc USA, Inc. v. Microsoft Corp., 632 F.3d 1292, 1320 (Fed. Cir. 2011); Lucent Techs., Inc. v. Gateway, Inc., 580 F.3d 1301, 1336 (Fed. Cir. 2009); Ericsson, Inc. v. D-Link Sys., Inc., 773 F.3d 1201, 1226 (Fed. Cir. 2014) (citing VirnetX, Inc. v. Cisco Sys., Inc., 767 F.3d 1308 (Fed. Cir. 2014) ("[W]here multi-component products are involved, the governing rule is that the ultimate combination of royalty base and royalty rate must reflect the value attributable to the infringing features of the product, and no more.")).

[287] *See* NAT'L RESEARCH COUNCIL OF THE NAT'L ACADS., *supra* note 36.

[288] *See* Commonwealth Sci. & Indus. Research Org. v. Cisco Sys., Inc., 809 F.3d 1295 (Fed. Cir. 2015); *Ericsson*, 773 F.3d at 1229–35; Microsoft Corp. v. Motorola, Inc., No. C10-

5. *Design Patents*

Design patents, which afford fifteen years of protection for "new, original and ornamental designs for an article of manufacture,"[289] have come into play in some network technology markets. As with copyright and trademark protection, design patents do not extend to the functionality of useful articles.[290] Only utility patent protection can protect such elements.

Separating ornamental from functional features has proven difficult. The Federal Circuit will invalidate a design patent only if the claimed design is *dictated solely* by the function of the article of manufacture.[291] Some decisions have applied a looser balancing test: asking whether a design is "primarily functional."[292]

The difficulty lies in the fact that functionality is often intertwined with ornamentality, especially in minimalist designs that merge form with function. Furthermore, compilations of design features can themselves be functional. Some Federal Circuit decisions address this challenge by dissecting the claimed design through a process that aligns with copyright law's treatment of the idea-expression dichotomy.[293] Thus, if a claimed design contains "both functional and ornamental features, the patentee must show that the perceived similarity is based on the ornamental features of the design."[294] The courts "factor[] out the functional aspects of [the claimed design] as part of its claim construction." [295] This approach, however, is in tension with the Federal

1823JLR, 2013 WL 2111217 (W.D. Wash. Apr. 25, 2013); *In re* Innovatio IP Ventures, LLC, No. 11 C 9308, 2013 WL 5593609 (N.D. Ill. Oct. 3, 2013).

[289] 35 U.S.C. § 171 (2018).

[290] *See* Christopher Buccafusco & Mark A. Lemley, *Functionality Screens*, 103 VA. L. REV. 1293, 1295 (2017); *see* Jason J. Du Mont & Mark D. Janis, *Functionality in Design Protection Systems*, 19 J. INTELL. PROP. L. 261, 271 (2012).

[291] *See* Best Lock Corp. v. Ilco Unican Corp., 94 F.3d 1563, 1566 (Fed. Cir. 1996).

[292] L.A. Gear, Inc. v. Thom McAn Shoe Co., 988 F.2d 1117, 1123 (Fed. Cir. 1993)

> [T]he utility of each of the various elements that comprise the design is not the relevant inquiry with respect to a design patent. In determining whether a design is primarily functional or primarily ornamental the claimed design is viewed in its entirety, for the ultimate question is not the functional or decorative aspect of each separate feature, but the overall appearance of the article, in determining whether the claimed design is dictated by the utilitarian purpose of the article.

Lee v. Dayton-Hudson Corp., 838 F.2d 1186, 1188 (Fed. Cir. 1988).

[293] *See supra* Section V.B.2.a).

[294] *See* OddzOn Prods., Inc. v. Just Toys, Inc., 122 F.3d 1396, 1405 (Fed. Cir. 1997).

[295] *See* Richardson v. Stanley Works, Inc., 597 F.3d 1288, 1293 (Fed. Cir. 2010).

Circuit's holding that claimed designs should be evaluated as a whole.[296] Some decisions have suggested that courts can surmount the separability challenge by considering

> whether the protected design represents the best design; whether alternative designs would adversely affect the utility of the specified article; whether there are any concomitant utility patents; whether the advertising touts particular features of the design as having specific utility; and whether there are any elements in the design or an overall appearance clearly not dictated by function.[297]

This standard parallels an earlier formulation of trademark law's functionality doctrine.[298] This approach reflects a concern with design patents preempting competition. A design is functional if there are no alternative designs that accomplish a function equally well.[299]

The Federal Circuit applied these principles in a case involving interoperability. In *Best Lock Corp. v. Ilco Unican Corp.*,[300] Best Lock claimed an unusual profile for a key blade blank, the form used for manufacturing (cutting) keys. Ilco distributed key blanks with that key blade shape. The Federal Circuit found that function alone dictated the key blade design because "no alternative blank key blade would fit the corresponding lock."[301]

The integration of form and function in many product markets has brought design patents into play in some network technology markets. Most notably, Apple successfully asserted design patents covering the rounded

[296] *See* Egyptian Goddess, Inc. v. Swisa, Inc., 543 F.3d 665, 677 (Fed. Cir. 2008) (en banc) (rejecting focusing on a design's "point of novelty"); Crocs, Inc. v. Int'l Trade Comm'n, 598 F.3d 1294, 1302–03 (Fed. Cir. 2010).

[297] PHG Techs., LLC v. St. John Cos., 469 F.3d 1361, 1366 (Fed. Cir. 2006); *see* Berry Sterling Corp. v. Pescor Plastics, Inc., 122 F.3d 1452, 1456 (Fed. Cir. 1997).

[298] *See In re* Morton-Norwich Prods., Inc., 671 F.2d 1332, 1340–41 (C.C.P.A. 1982); *see also* Amini Innovation Corp. v. Anthony Cal., Inc., 439 F.3d 1365, 1371 (Fed. Cir. 2006) (stating that an "aspect" of a patented design is functional "if it is essential to the use or purpose of the article or if it affects the cost or quality of the article," which is a trademark functionality standard articulated in *Inwood Labs., Inc.*, 456 U.S. at 850, n.10, discussed in *supra* Section V.C).

[299] *See, e.g.*, Rosco, Inc. v. Mirror Lite Co., 304 F.3d 1373, 1378 (Fed. Cir. 2002) (reasoning that "if other designs could produce the same or similar functional capabilities, the design of the article in question is likely ornamental, not functional"); Seiko Epson Corp. v. Nu-Kote Int'l, Inc., 190 F.3d 1360, 1368 (Fed. Cir. 1999) (explaining that "the design must not be governed solely by function, i.e., that this is not the only possible form of the article that could perform its function"); L.A. Gear, Inc. v. Thom McAn Shoe Co., 988 F.2d 1117, 1123 (Fed. Cir. 1993) ("When there are several ways to achieve the function of an article of manufacture, the design of the article is more likely to serve a primarily ornamental purpose.").

[300] 94 F.3d 1563.

[301] *Id.* at 1566.

rectangular shape of mobile communications devices against Samsung.[302] Technology companies and designers have also obtained design patents on virtual designs, patents that cover the designs of graphical user interfaces for smartphones, tablets, and other products, as well as the designs of icons or other artifacts of various virtual environments.[303]

The strong monetary remedies available for design patents further encourage seeking design patents to protect features of network technologies. The Patent Act provides for recovery of the "total profit" on the sale of "any article of manufacture to which [a protected design] has been applied."[304] Although the Supreme Court held that the term "article of manufacture" encompasses both a product and a component of that product,[305] the apportionment of damages in design patent cases is uncertain.

VI. INTERPLAY OF INTELLECTUAL PROPERTY PROTECTION AND COMPETITION POLICY IN NETWORK INDUSTRIES

The Sherman Antitrust Act prohibits contracts in restraint of trade and monopolization or attempts to monopolize markets.[306] The courts have long recognized that patent and copyright protections—government-authorized rights to exclude others from using protected technologies and copying works of authorship—function as limited exceptions to antitrust liability.[307] Yet,

[302] *See* Apple Inc. v. Samsung Elecs. Co., 786 F.3d 983 (Fed. Cir. 2015) (holding that iPhone and iPad designs were functional (and hence unprotectable under trademark law) but not functional under design patent law).

[303] *See* Jason J. Du Mont & Mark D. Janis, *Virtual Designs*, 17 STAN. TECH. L. REV. 107 (2013).

[304] 35 U.S.C. § 289 (2018).

[305] *See* Samsung Elecs. v. Apple Inc., 137 S. Ct. 429 (2016).

[306] *See* 15 U.S.C. §§ 1–2 (2018).

[307] *See* FTC v. Actavis, 133 S. Ct. 2223, 2238 (2103) (Roberts, C.J., dissenting) (observing that a patent "provides an exception to antitrust law, and the scope of the patent—i.e., the rights conferred by the patent—forms the zone within which the patent holder may operate without facing antitrust liability"); Simpson v. Union Oil Co. 377 U.S. 13, 24 ("The patent laws which give a 17-year monopoly on 'making, using, or selling the invention' are *in pari materia* with the antitrust laws and modify them *pro tanto*."); SCM Corp. v. Xerox Corp., 645 F.2d 1195, 1209 (2d Cir. 1981) (imposing antitrust liability for patentee's refusal to license on lawfully acquired patent "would severely trample upon the incentives provided by our patent laws and thus undermine the entire patent system"); Image Tech. Servs. v. Eastman Kodak Co., 125 F.3d 1195, 1217–18 (9th Cir. 1997) (recognizing that "[w]ithout bounds, claims based on unilateral [refusals to deal]" by patent and copyright holders "will proliferate" and "[t]he cost of such suits will reduce [their] 'incentive . . . to risk the often enormous costs in terms of time,

intellectual property rights can concentrate economic power in ways that undermine competition, leading courts and antitrust enforcement agencies to develop more nuanced and complementary views of the interplay between intellectual property protection and antitrust liability.[308] This is especially true in network technology markets, where positive feedback effects often lead to strong and durable monopolies. At the same time, high concentration can promote desirable network effects.

These considerations ameliorate and complicate the interplay of intellectual property protection and competition policy. Section VI.A explores limitations on improper leveraging of intellectual property rights that arise in private enforcement of intellectual property and contracts. Section VI.B examines public enforcement of antitrust law and competition policy in network markets.

research and development' ") (quoting Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 480 (1974)).

The patent and copyright misuse doctrines, however, serve as exceptions to the limited exception. *See* Motion Picture Patents Co. v. Universal Film Mfg. Co., 243 U.S. 502 (1917) (recognizing the patent misuse doctrine); Carbice Corp. of Am. v. Am. Patents Dev. Corp., 283 U.S. 27, supplemented, 283 U.S. 420 (1931) (holding that the patentee could not condition the right to use the patented invention on the purchase of unpatented materials, thereby establishing the staple article of commerce doctrine); Lasercomb Am., Inc. v. Reynolds, 911 F.2d 970 (4th Cir. 1990) (recognizing a copyright misuse doctrine).

[308] *See* Atari Games Corp. v. Nintendo of Am., Inc., 897 F.2d 1572, 1576 (Fed. Cir. 1990) ("[T]he aims and objectives of patent and antitrust laws may seem, at first glance, wholly at odds. However, the two bodies of law are actually complementary, as both are aimed at encouraging innovation, industry, and competition."); Marina Lao, *Unilateral Refusals to Sell or License Intellectual Property and the Antitrust Duty to Deal*, 9 CORNELL J.L. & PUB. POL'Y 193, 193 (1999) ("Courts and academics alike considered intellectual property rights as exceptions to the antitrust law that must be narrowly construed."); U.S. DEP'T OF JUSTICE & FED. TRADE COMM'N, ANTITRUST GUIDELINES FOR THE LICENSING OF INTELLECTUAL PROPERTY (1995), *reprinted in* 4 TRADE REG. REP. (CCH) ¶ 13,132

> The intellectual property laws and the antitrust laws share the common purpose of promoting innovation and enhancing consumer welfare. The intellectual property laws provide incentives for innovation and its dissemination and commercialization by establishing enforceable property rights for the creators of new and useful products, more efficient processes, and original works of expression . . . . The antitrust laws promote innovation and consumer welfare by prohibiting certain actions that may harm competition with respect to either existing ways or new ways of serving consumers.

Willard K. Tom & Joshua A. Newberg, *Antitrust and Intellectual Property: From Separate Spheres to Unified Field*, 66 ANTITRUST L.J. 167 (1997).

A. Private Enforcement

Courts have recognized limits on the exercise of patent and copyright protection that apply with special force in network industries. As we have already seen, various internal intellectual property doctrines—such as copyright's fair use doctrine and the use of equitable balancing in dispensing remedies—bring competition policy concerns into intellectual property law. In addition, courts have developed equitable and contract-based defenses to prevent anti-competitive abuses of intellectual property rights.

1. *Misuse Doctrines*

Drawing on tort law's "unclean hands" doctrine, courts developed the patent misuse doctrines as a common law equitable defense to an infringement claim.[309] Unlike the purely equitable defense of "unclean hands," the misuse doctrines apply to suits for damages as well as equitable relief. The misuse doctrine bars patent owners from expanding the scope or term of the intellectual property right through licensing restrictions. The Supreme Court prevented Thomas Edison from leveraging a patent on motion picture projectors to control what films could be exhibited using that projector.[310] The doctrine bars enforcement of the patent until the anti-competitive effects of the restriction have been purged.

The patent misuse doctrine applies whether or not an antitrust violation has been established. The expansion of the patent misuse doctrine in the 1940s led Congress to exclude contributory patent infringement claims from the ambit of patent misuse in the 1952 Patent Act.[311] Nonetheless, the uncertain scope and severe remedy of patent misuse continued to generate criticism, especially as economists and courts came to question categorical antitrust prohibitions in favor of rule of reason balancing.[312]

[309] *See* Christina Bohannan, *IP Misuse as Foreclosure*, 96 IOWA L. REV. 475, 484–86, 515 (2011).

[310] *See* Motion Picture Patents Co. v. Universal Film Mfg. Co., 243 U.S. 502 (1917) (refusing to enforce a licensing provision restricting use of the machine to motion pictures licensed by Edison's film company); *see Carbice Corp.*, 283 U.S.; *see also* Morton Salt Co. v. G.S. Suppiger Co., 314 U.S. 488 (1942).

[311] *See* 35 U.S.C. § 271(d)(1)–(3) (2018).

[312] *See* Herbert Hovenkamp, *Antitrust and the Patent System: A Reexamination*, 76 OHIO ST. L.J. 467, 485, 497 (2015); Bohannan, *supra* note 309, at 490–95; *see generally* Mark A. Lemley, *The Economic Irrationality of the Patent Misuse Doctrine*, 78 CALIF. L. REV. 1599 (1990); Herbert J. Hovenkamp, *The Rule of Reason and the Scope of the Patent*, 52 SAN DIEGO L. REV. 515 (2015) [hereinafter Hovenkamp, *The Rule of Reason*]; Robin C. Feldman, *The Insufficiency of Antitrust Analysis for Patent Misuse*, 55 HASTINGS L.J. 399, 422–23 (2003); *see generally* Robert P. Merges,

Congress amended the Patent Act in 1988 to codify two additional patent misuse limitations.[313] Congress insulated refusals to license any rights to a patent from charges of patent misuse.[314] This provision, however, can effectively be side-stepped through contract, such as a FRAND commitment. In that circumstance, the third-party beneficiary of the SSO agreement has a breach of contract action for failure to license standard essential patents on FRAND terms.[315] Furthermore, Congress barred application of the patent misuse doctrine to tying arrangements unless the patentee has market power in the relevant market for the patent or has patented tying a product on which the license or sale is conditioned,[316] thereby bringing patent misuse more closely in line with antitrust liability.[317]

The courts have struggled to disentangle patent misuse doctrine from antitrust analysis.[318] Although the two fields share common concerns, the misuse doctrine has sought to promote intellectual property policies of encouraging innovation, freedom to operate outside of intellectual property protections, and access to the public domain even when objectionable practices do not violate antitrust law.[319]

In a case echoing the *Motion Picture Patents* case, a music copyright licensor sought to require theaters to obtain a performance license before they even knew what music would be incorporated into the films they would show.[320] The district court found that such a license agreement improperly asserted control over all films and hence constituted copyright misuse. As a result, the court barred enforcement against the theater owner.

As the emergence of computer software brought copyright more directly into play in innovation markets, the copyright misuse doctrine has come into wider use. In *Lasercomb America, Inc. v. Reynolds*,[321] a software copyright licensor

Reflections on Current Legislation Affecting Patent Misuse, 70 J. PAT. & TRADEMARK OFF. SOC'Y 793, 793 (1988) (noting that "the often very limited (or 'thin') markets for patented technology make it difficult to apply antitrust law's consumer-demand definition of the relevant market").

[313] Patent Misuse Reform Act of 1988, Pub. L. No. 100-703, 102 Stat. 4674 (1988).

[314] *See* 35 U.S.C. § 271(d)(4).

[315] *See infra* Section VI.A.4.

[316] 35 U.S.C. § 271(d)(5).

[317] *See generally* Hovenkamp, *The Rule of Reason*, *supra* note 312, at 515.

[318] *See* Feldman, *supra* note 312; Bohannan, *supra* note 312, at 490–95; Janice M. Mueller, *Patent Misuse Through the Capture of Industry Standards*, 17 BERKELEY TECH. L.J. 623, 653–57 (2002); *see generally* Robert Pitofsky, *Challenges of the New Economy: Issues at the Intersection of Antitrust and Intellectual Property*, 68 ANTITRUST L.J. 913 (2001).

[319] *See* Bohannan, *supra* note 309, at 505.

[320] *See* M. Witmark & Sons v. Jensen, 80 F. Supp. 843 (D. Minn. 1948), *appeal dismissed*, *sub nom.* M. Witmark & Sons v. Berger Amusement Co., 177 F.2d 515 (8th Cir. 1949).

[321] 911 F.2d 970 (4th Cir. 1990).

prohibited licensees from "writing, developing, producing or selling computer assisted die making software, directly or indirectly without Lasercomb's prior written consent" for a term of ninety-nine years. Drawing on the principles underlying intellectual property protection as well as patent misuse jurisprudence, the court determined that copyright misuse is a valid defense and barred Lasercomb's infringement action. In this case, the licensor sought to foreclose competition in computer software innovation.

The copyright misuse doctrine has since been raised in a variety of settings, including tying arrangements, anticompetitive clauses in licensing agreements, mandatory blanket licenses, and refusals to license, but it remains murky.[322] Several cases in which copyright misuse has been found to be viable involve network effects. For example, the Ninth Circuit held that the licensing terms of the American Medical Association's (AMA) Physician's Current Procedural Terminology to the Health Care Financing Administration gave the AMA a substantial and unfair advantage over its competitors and hence constituted copyright misuse.[323] The Fifth Circuit in *DSC Communications Corp. v. DGI Technologies, Inc.* held that the copyright misuse doctrine might be viable to defend assertions of copyright protection over interoperable features of computer software.[324] The Seventh Circuit in *Assessment Technologies of WI, LLC v. WIREdata, Inc.* held that licensing restrictions on a tax assessment database to control access to public domain data inputted by public tax assessors could constitute copyright misuse.[325]

As with the patent misuse doctrine, the interplay of copyright misuse doctrine and antitrust liability remains unclear. Courts applying the copyright misuse doctrine generally evaluate whether the conduct thwarts the underlying policies of copyright law. Some courts, however, mistakenly view the doctrine as co-extensive with antitrust law.[326]

2. *The Principle of Exhaustion*

With some resemblance to misuse, the long-standing common law doctrine of exhaustion (also known as first-sale) preserves the public interest in free competition by limiting the ability of IP owners to control secondary

[322] *See* Jonas P. Herrell, *The Copyright Misuse Doctrine's Role in Open and Closed Technology Platforms*, 26 BERKELEY TECH. L.J. 441, 466 (2011); *see also* Aaron Xavier Fellmeth, *Copyright Misuse and the Limits of the Intellectual Property Monopoly*, 6 J. INTELL. PROP. L. 1, 24 (1998).

[323] *See* Practice Mgmt. Info. Corp. v. Am. Med. Ass'n, 121 F.3d 516 (1995), *amended by*, 133 F.3d 1140 (9th Cir. 1997).

[324] 81 F.3d 597 (5th Cir. 1996).

[325] 350 F.3d 640, 647 (7th Cir. 2003).

[326] *See* Fellmeth, *supra* note 322, at 22–23.

markets for patented products and copyrights works. [327] The idea is straightforward: the first authorized sale of the patented or copyrighted product exhausts the monopolistic power given to the owner as a reward for its efforts and contribution to society.

Following the first sale, the owner can no longer control the manner in which the product is sold or used in secondary markets, either by downstream purchasers or subsequent sellers. This limitation on monopoly, which traces back to the common law's general hostility towards restraints on alienation, fosters competition in secondary markets for innovative and creative works.

Recently, in a much-debated decision involving Lexmark cartridges, the Supreme Court emphasized the destructive anti-competitive effects post-sale restrictions that "run with" the product and exhibit servitude-like features have on free commerce.[328] Refusing to allow patentees to "sputter" "the smooth flow of commerce,"[329] the Court held that the principle of exhaustion prevents the enforcement of contractual post-sale restrictions through patent law.[330] The patentee can impose contractual restrictions on secondary markets but cannot use patent law to control how the product is being used or sold downstream after the point of the first sale. The *Lexmark* decision reaffirmed the vital role that IP limiting doctrines such as preemption, exhaustion, and misuse play in limiting IP owners' ability to hinder downstream innovation thorough over-reaching, often boilerplate, contractual language.[331]

3. *Ambush of Standard-Setting Processes*

As highlighted in Section V.D.3, SSOs play a critical role in addressing the market failures surrounding network technologies. These organizations require companies that participate in standard setting processes to disclose relevant

[327] *See* Adams v. Burke, 84 U.S. 453 (1873) (holding that once patented products have been sold, the patent holder may not restrict the way those products are used or sold in secondary markets); Bobbs-Merrill Co. v. Straus, 210 U.S. 339 (1908) (confirming the "first sale" doctrine in copyright law). Congress codified the exhaustion principle in the Copyright Act of 1976. 17 U.S.C. § 109(a) (2018).

[328] *See* Impression Prods. v. Lexmark Int'l, Inc., 137 S. Ct. 1523, 1530–32 (2017); Molly Shaffer Van Houweling, *The New Servitudes*, 96 GEO. L.J. 921 (2008) (describing the burden of servitude on free commerce); *see also* Molly Shaffer Van Houweling, *Intellectual Property as Property*, *in* 1 RESEARCH HANDBOOK ON THE ECONOMICS OF INTELLECTUAL PROPERTY (Ben Depoorter & Peter S. Menell eds., 2019).

[329] *See Impression Prods.*, 137 S. Ct. at 1532.

[330] *See id.* at 1531–33.

[331] *See* Amit Elazari Bar-On, *Unconscionability 2.0 and the IP Boilerplate: A Revised Doctrine of Unconscionability for the Information Age*, 34 BERKELEY TECH. L.J. (forthcoming 2019); AARON PERZANOWSKI & JASON SCHULTZ, THE END OF OWNERSHIP: PERSONAL PROPERTY IN THE DIGITAL ECONOMY chs. 9–10 (2016).

information about actual and potential patents implicated by draft standards and commit to license such technologies on FRAND terms.[332] Given the dynamic nature of technological progress, such conditions can be open to interpretation.

The two most prominent cases alleging that SSO participants engaged in deceptive practices—one involving Rambus and the other involving Qualcomm—reached different conclusions. The Rambus litigation grew out of standards development for dynamic random-access memory (DRAM) chips, a memory technology that was widely adopted throughout the computer industry.[333] In 1990, Rambus sought patents on its architecture for such chips. Around that time, the Joint Electron Device Engineering Council (JEDEC) organized an open standard setting process with a broad range of industry participants including Rambus. As would be revealed in later litigation, Rambus used information gained at the meetings to amend its patent applications so that the standards would read on its patents. Rambus concealed these efforts and subsequently withdrew from JEDEC. After the JEDEC standards gained widespread acceptance in products, Rambus began making royalty demands from implementers and, beginning in 2000, brought a series of enforcement actions. The jury in one of the key cases found that Rambus committed fraud and breached its JEDEC obligations by failing to disclose its patents. The Federal Circuit reversed in a divided opinion, with the majority finding that the JEDEC policy statements were too vague to support a fraud finding.[334]

The Qualcomm litigation grew out of the development of the Joint Video Team (JVT) standard for video compression technology.[335] Qualcomm, a pioneer in semiconductor design for mobile communications devices and various other technologies, participated in the JVT standard setting process. It later sought to enforce several patents applicable to that standard against Broadcom. Broadcom successfully defended on the ground that Qualcomm had waived its rights to enforce the patent as a result of its failure to disclose the patents as part of the standard setting process. The court barred Qualcomm from enforcing the patents at issue against any products implementing the pertinent JVT standard.

[332] *See* Mark R. Patterson, *Leveraging Information About Patents: Settlements, Portfolios, and Holdups*, 50 HOUS. L. REV. 483, 513–21 (2012).

[333] *See* Rambus Inc. v. Infineon Techs. AG, 318 F.3d 1081 (Fed. Cir. 2003).

[334] *See id.* at 1098.

[335] *See* Qualcomm Inc. v. Broadcom Corp., 548 F.3d 1004 (Fed. Cir. 2008).

The different results in these litigations reflect several factors. SSO policies in the early 1990s varied in how clearly they set forth disclosure requirements. Furthermore, defendants in these cases faced a variety of complex evidentiary requirements and heightened pleading standards to equitable defenses such as laches, waiver, actual or implied license, equitable estoppel, and fraud.[336] The Rambus controversy led to public enforcement actions in the United States and Europe that are discussed in Section VI.B.2.

4. *Breach of Contract for Failure to License SEPs on FRAND Terms*

As explored earlier,[337] SSOs typically require companies participating in standard setting processes to commit to license standard essential patents on FRAND terms. The litigation between Microsoft and Motorola established key principles regarding the determination of FRAND licensing terms and provided for the award of contract damages for breach of the FRAND commitment. Motorola participated in standard setting processes governed by the IEEE and the International Telecommunication Union (ITU) establishing Wi-Fi (802.11) and video compression (H.264) standards.[338] As part of its participation in these processes, Motorola agreed to license its patents that are essential to those standards on reasonable and non-discriminatory (RAND)[339] terms.

The controversy began in October 2010 when Microsoft filed actions in the ITC and the Western District of Washington alleging that Motorola was infringing several Microsoft smartphone patents.[340] Those filings immediately led to settlement negotiations involving cross-licensing of patents between the two companies. Later that month, Motorola sent letters to Microsoft requesting royalties equal to 2.25% of Microsoft's sales revenues from Windows and Xbox products incorporating the standards. Microsoft declined and immediately filed suit in the Western District of Washington alleging that Motorola's offer breached its RAND commitment. Microsoft asserted that it was a third-party beneficiary of the SSO agreements. Thereupon Motorola filed patent-enforcement suits with the ITC, seeking an exclusion order against importing Microsoft's Xbox products into the United States, and with a German court, seeking an injunction against sales of Microsoft's H.264-

[336] *See* Sean Royall, Amanda Tessar & Adam Di Vincenzo, *Deterring "Patent Ambush" in Standard Setting: Lessons from* Rambus *and* Qualcomm, 23 ANTITRUST 34 (2009).

[337] *See supra* Section V.D.3–4.

[338] *See* Microsoft Corp. v. Motorola, Inc., No. C10-1823JLR., 2013 WL 2111217 (W.D. Wash. Apr. 25, 2013) at *5, *8, *21, *49, *53.

[339] FRAND and RAND are used interchangeably in the technology industries. FRAND is the more common usage.

[340] *See Microsoft Corp.*, 2013 WL 2111217.

compliant products. The German action threatened all of Microsoft's Windows and Xbox European sales because its distribution center was located in Germany. As a result, Microsoft immediately relocated its distribution center to the Netherlands at substantial cost.

Judge Robart adapted the *Georgia-Pacific* reasonable royalty framework to the FRAND context.[341] In so doing, he set forth the following principles:

> (1) A RAND royalty should be set at a level consistent with the SSOs' goal of promoting widespread adoption of their standards[;]
>
> (2) A proper methodology should] recognize and seek to mitigate the risk of patent hold-up that RAND commitments are intended to avoid[;]
>
> (3) [A] proper methodology for determining a RAND royalty should address the risk of royalty stacking by considering the aggregate royalties that would apply if other SEP holders made royalty demands of the implementer[;]
>
> (4) At the same time, a RAND royalty should be set with the understanding that SSOs include technology intended to create valuable standards[, which requires that] the RAND commitment [] guarantee that holders of valuable intellectual property will receive reasonable royalties on that property[; and]
>
> (5) From an economic perspective, a RAND commitment should be interpreted to limit a patent holder to a reasonable royalty on the economic value of its patented technology itself, apart from the value associated with incorporation of the patented technology into the standard.[342]

Applying these economic guideposts, Judge Robart concluded that the reasonable royalty should be approximately 1/100th of Motorola's 2.25% license offer. Motorola's patents constituted less than 10% of the Wi-Fi 802.11 pool and none was shown to be of special importance.[343] Judge Robart noted that if each of ninety-two companies that owned SEPs for the 802.11 and H.264 standards demanded a royalty rate comparable to Motorola's offer, the sum of the royalties would exceed the selling price of the Xbox.[344] In a later proceeding, a jury awarded Microsoft $14.52 million ($11.49 million for

[341] *See Microsoft Corp.*, 2013 WL 2111217.

[342] *Id.* at *12; *see generally* William H. Page, *Judging Monopolistic Pricing: F/RAND and Antitrust Injury*, 22 TEX. INTELL. PROP. L.J. 181 (2014).

[343] *See Microsoft Corp.*, 2013 WL 2111217, at *101.

[344] *Id.* at *52, *72–73.

relocating its distribution center and $3.03 million in attorneys' fees and litigation costs) for Motorola's breach of its contractual commitment to license its standard essential patents on RAND terms.[345]

The Ninth Circuit upheld these determinations, expressly recognizing the key role of RAND commitments in "mitigat[ing] the risk that a SEP holder will extract more than the fair value of its patented technology . . . . Under these agreements, a SEP holder cannot refuse a license to a manufacturer who commits to paying the RAND rate."[346]

5. *Private Antitrust Liability*

Section 4 of the Clayton Antitrust Act of 1914 authorizes recovery of damages by "any person . . . injured in his business or property by reason of anything forbidden in the antitrust laws," including section 7 of the Sherman Act.[347] Companies have used this private right of action to combat anti-competitive practices in network industries. Three issues are particularly relevant to network technology markets: (i) refusal to license patented technologies; (ii) patent thickets; and (iii) leveraging of monopoly power.

a) Refusals to License Patented Technologies and Copyright-Protected Works

As noted above, the Patent Act grants patentees the exclusive right to use patented technologies. Congress reinforced that power by expressly providing that a refusal to license a patent cannot be the basis for a patent misuse defense. The courts are divided over whether a refusal to license patented technology can constitute an antitrust violation. The so-called "essential facilities" doctrine, which holds that "an owner of a crucial input cannot deny access if a firm seeking access cannot practicably obtain the input elsewhere"[348] has lost favor among commentators[349] and the courts.[350] The courts focus on whether

[345] *See id.*, at *101.

[346] Microsoft Corp. v. Motorola, Inc., 795 F.3d 1024, 1031 (9th Cir. 2015) (affirming Judge Robart's decision).

[347] 15 U.S.C. § 15 (1914).

[348] James R. Ratner, *Should There Be an Essential Facility Doctrine?*, 21 U.C. DAVIS L. REV. 327, 330 (1988); Otter Tail Power Co. v. United States, 410 U.S. 366, 371, 377 (1973) (holding that a single firm's refusal to deal with other firms that denied them access to a facility essential to engaging in business violates antitrust law).

[349] *See generally* Phillip Areeda, *Essential Facilities: An Epithet in Need of Limiting Principles*, 58 ANTITRUST L.J. 841 (1989); *see also* Ratner, *supra* note 348.

[350] *See* Eastman Kodak Co. v. Image Tech. Servs., Inc., 504 U.S. 451 (1992) (declining to apply essential facilities doctrine); Aspen Skiing Co. v. Aspen Highlands Skiing Corp., 472 U.S. 585 (1985) (same).

the defendant has a legitimate business justification for its conduct.[351] The key modern cases involve the control of service and replacement part aftermarkets.

In the late 1980s, a variety of companies that had entered the market to service Kodak photocopiers found themselves cut off from replacement parts.[352] Kodak ended its practice of licensing and selling replacement parts to competing service companies and required that its original equipment manufacturers not sell parts to independent service operators. The independent service organizations brought suit, claiming that Kodak unlawfully tied the sale of service for Kodak machines with the sale of parts in violation of section 1 of the Sherman Act, and monopolized or attempted to monopolize the sale of service for Kodak machines in violation of section 2 of the Sherman Act. Kodak defended in part on its intellectual property rights.

While recognizing that patent and copyright owners have exclusive rights to their protected works, the Ninth Circuit nonetheless held that these laws only afford the intellectual property owner a rebuttable "presumptively valid business justification" for consumer harm.[353] The court upheld a jury verdict finding Kodak liable for monopolizing or attempting to monopolize the service aftermarket for Kodak copiers.[354]

In an analogous case involving Xerox's refusal to sell patented parts and copyrighted manuals and to license copyrighted software, the Federal Circuit declined to follow the Ninth Circuit's analysis.[355] The Federal Circuit limited its focus to whether Xerox's refusal to sell its patented parts exceeded the scope of the patent grant. Finding that it did not, "Xerox was under no obligation to sell or license its patented parts and did not violate the antitrust laws by refusing to do so."[356] The court ruled that so long as a patent infringement suit would not have been objectively baseless, the patentee's motivations for asserting its statutory right to exclude are immaterial. Similarly, the Federal Circuit further held that so long as Xerox's copyrights were not "obtained by unlawful means or were used to gain monopoly power beyond" the statutory grant, then "Xerox's refusal to sell or license its copyrighted

[351] *See, e.g.*, Jonathan B. Baker, *Promoting Innovation Competition Through the* Aspen/Kodak *Rule*, 7 GEO. MASON L. REV. 495, 503–08 (1999).

[352] *See* Image Tech. Servs., Inc. v. Eastman Kodak Co., 125 F.3d 1195, 1200–01 (9th Cir. 1997).

[353] *See id.* at 1218 (quoting Data General Corp. v. Grumman Sys. Support Corp., 436 F.3d 1147, 1187 (1st Cir. 1994)).

[354] *See id.* at 1228.

[355] *In re* Indep. Serv. Org. Antitrust Litig., 203 F.3d 1322 (Fed. Cir. 2000); *see* Intergraph Corp. v. Intel Corp., 195 F.3d 1346 (Fed. Cir. 1999).

[356] *Indep. Serv. Org.*, 203 F.3d at 1328.

works was squarely within the rights granted by Congress to the copyright holder and did not constitute a violation of the antitrust laws."[357]

b) Patent Thickets

In a related vein, competitors have sought to challenge the accumulation of a broad portfolio of patents on antitrust grounds. Accumulation and pooling of patents can broaden the effective scope and reduce the uncertainty surrounding inventions, thereby enhancing appropriability.[358] Nonetheless, strong and broad patent portfolios can discourage innovation and entry by potential competitors.[359]

The leading case involved Xerox Corporation, which built a portfolio of over 1,000 patents relating to its plain paper copying technology. Xerox only used 35 to 40 percent of those patents in actual Xerox products,[360] relying on the balance to erect a defensive thicket around its photocopier technology.[361] Xerox refused to grant licenses for plain paper copying, although it did grant some licenses for other fields, including coated paper copiers. SCM Corporation, which had licensed some of Xerox's patents for coated paper copies, filed an antitrust claim against Xerox alleging that "Xerox's acquisition of its patents and subsequent exercise of the exclusionary power in them violated the antitrust laws and injured SCM."[362] SCM asserted that Xerox's patent accumulation strategy was intended to forestall competition, as reflected in its failure to use many of its patents.[363] It further argued that "Xerox's acquisition of its patents and subsequent exercise of the exclusionary power in them violated the antitrust laws and injured SCM."[364] The Second Circuit acknowledged that "tension between the objectives of preserving economic

[357] *Id.* at 1329.

[358] *See* Gideon Parchomovsky & R. Polk Wagner, *Patent Portfolios*, 54 U. PA. L. REV. 1, 32–41 (2005).

[359] *See generally* Bronwyn H. Hall, Christian Helmers & Georg von Graevenitz, *Technology Entry in the Presence of Patent Thickets* (Inst. for Fiscal Studies, Working Paper No. 16-02, 2016); Herbert Hovenkamp, *Antitrust and the Movement of Technology*, 19 GEO. MASON L. REV. 1119, 1130 (2012) (discussing the costs of defending against many patents of ambiguous scope); Daniel L. Rubinfeld & Robert Maness, *The Strategic Use of Patents: Implications for Antitrust*, *in* 1 ANTITRUST, PATENTS, AND COPYRIGHT: EU AND US PERSPECTIVES 85 (François Lévêque & Howard A. Shelanski eds., 2005).

[360] *See* Gerald Sobel, *The Antitrust Interface with Patents and Innovation: Acquisition of Patents, Improvement Patents and Grant-Backs, Non-Use, Fraud on the Patent Office, Development of New Products and Joint Research*, 53 ANTITRUST L.J. 681 (1984).

[361] *See* Kurt M. Saunders, *Patent Nonuse and the Role of Public Interest as a Deterrent to Technology Suppression*, 15 HARV. J.L. & TECH. 389, 393 (2002).

[362] SCM Corp. v. Xerox Corp., 645 F.2d 1195, 1203 (2d Cir. 1981).

[363] *See id.* at 1202–03.

[364] *Id.* at 1203.

incentives to enhance competition while at the same time trying to contain the power a successful competitor acquires is heightened tremendously when the patent laws come into play," emphasizing that the Xerox case "demonstrate[s that] the acquisition of a patent can create the potential for tremendous market power."[365] Nonetheless, the court ultimately ruled that "where a patent has been lawfully acquired, subsequent conduct permissible under the patent laws cannot trigger any liability under the antitrust laws."[366]

The anticompetitive concerns relating to patent thickets are exacerbated by the ambiguity of many software patent claims.[367] Cross-licensing and patent pools can, however, alleviate concerns about patent thickets.[368] Economists generally believe that the inclusion of complementary and potentially blocking patents in a patent pool promotes competition by reducing the transaction costs and promoting licensing.[369]

c) Improper Leveraging of Market Power

In the mid-1990s, Microsoft Corporation held a dominant position in the desktop software marketplace just as the Internet emerged as an economic platform. Sun Microsystems's Java programming language for websites was rapidly gaining salience as a technology for easily transforming static webpages into engaging, animated, interactive websites. After failing to develop its own web development package, Microsoft entered into a Technology License and Distribution Agreement (TLDA) with Sun that allowed Microsoft to use, modify, and adapt Java technology in developing MS Internet Explorer 4.0 and other software products.[370] To safeguard Sun's WORA interoperability

[365] *Id.* at 1205.

[366] *Id.* at 1206.

[367] *See generally* James Bessen, *Patent Thickets: Strategic Patenting of Complex Technologies* (Research on Innovation, Boston Univ. School of Law, Working Paper No. 0401, 2004).

[368] *See* Jonathan M. Barnett, *The Anti-Commons Revisited*, 29 HARV. J.L. & TECH. 127 (2015); Jonathan M. Barnett, *From Patent Thickets to Patent Networks: The Legal Infrastructure of the Digital Economy*, 55 JURIMETRICS J. 1 (2014); Richard J. Gilbert, *Antitrust for Patent Pools: A Century of Policy Evolution*, 2004 STAN. TECH. L. REV. 3 (2004); Carl Shapiro, *Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting*, *in* 1 INNOVATION POL'Y & ECON. 119 (2000); Robert P. Merges, *Institutions for Intellectual Property Transactions: The Case of Patent Pools*, *in* 1 INNOVATION POLICY FOR THE KNOWLEDGE OF SOCIETY (Rochelle Cooper Dreyfuss, Diane Leenheer Zimmerman & Harry First eds., 2001).

[369] *See* Shapiro, *supra* note 368, at 144.

[370] *See* Sun Microsystems, Inc. v. Microsoft Corp., 21 F. Supp. 2d 1109, 1113–14 (N.D. Cal. 1998).

principle, the TLDA required that Microsoft adhere to Java's standardized application environment and compliance tests.[371]

Microsoft's deployment of its own version of Java, compatible only with other Microsoft products in violation of the WORA principle, threatened Sun's Java development strategy. In October 1997, Sun sued Microsoft for breach of contract, trademark infringement, copyright infringement, false advertising, and unfair competition.[372] In early 2002, Microsoft agreed to pay Sun twenty million dollars and was permanently prohibited from using "Java compatible" trademarks on its products.[373] The following year, Sun brought an antitrust and patent infringement action against Microsoft resulting in an award of over one billion dollars.[374]

B. PUBLIC ENFORCEMENT

Federal and state antitrust authorities have long played substantial roles in policing network market competition. The U.S. Department of Justice's filing of an antitrust action against IBM in 1969 reshaped the competitive landscape of the computer hardware industry and paved the way for a vibrant software industry.[375] At the time, IBM bundled software and services into the cost of leasing use of its hardware, limiting competitors' ability to charge for software development and products.[376] Immediately following the filing of the enforcement action, IBM unbundled software and services from its hardware sales thereby opening up markets for software products.[377] Although the

[371] *See id.* at 1114.

[372] *See* John Markoff, *Sun Sues Microsoft on Use of Java System*, N.Y. TIMES (Oct. 8, 1997), https://www.nytimes.com/1997/10/08/business/sun-sues-microsoft-on-use-of-java-system.html [https://perma.cc/Q6VK-55R8].

[373] *See* Stephen Shankland, *Sun, Microsoft Settle Java Suit*, CNET (Mar. 15, 2002), https://www.cnet.com/news/sun-microsoft-settle-java-suit/ [https://perma.cc/V5DY-KS3F].

[374] Scarlet Pruitt & Paul Roberts, *Update: Sun, Microsoft Settle Suit in Billion-Dollar Pact, Microsoft to Pay $700 Million for Antitrust Issues, $900 Million to Resolve Patent Dispute*, INFOWORLD (Apr. 2, 2004), http://www.infoworld.com/article/2667124/operating-systems/update--sun--microsoft-settle-suit-in-billion-dollar-pact.html [https://perma.cc/M9LD-6ZTL]; Stephen Shankland, *Sun Brings Antitrust Suit Against Microsoft, The Company Files a Private Antitrust Suit Against Microsoft Seeking Damages That Could Top $1 Billion*, CNET (July 20, 2002), https://www.cnet.com/news/sun-brings-antitrust-suit-against-microsoft-1/ [https://perma.cc/2RR4-VWKT].

[375] *See* Steven W. Usselman, *Unbundling IBM: Antitrust and the Incentives to Innovation in American Computing*, *in* THE CHALLENGE OF REMAINING INNOVATIVE 249, 249–50 (Sally H. Clarke, Naomi R. Lamoreaux & Steven W. Usselman eds., 2009).

[376] *See* Barack Richman & Steven Usselman, *Elhauge on Tying; Vindicated by History*, 49 TULSA L. REV. 689, 696–706 (2014).

[377] *See generally* Burton Grad, *A Personal Recollection: IBM's Unbundling of Software and Services*, 24 IEEE ANNALS HIST. COMPUTING 64 (2002).

antitrust case dragged on for more than a decade and was ultimately dropped, IBM's unbundling decision in conjunction with the emergence of mini and microcomputers markets revolutionized the computer industry. Similarly, the U.S. Department of Justice's filing of antitrust litigation against AT&T in 1974 led to the breakup of the largest corporation in the United States nearly a decade later and hastened the modern competitive and highly innovative telecommunications marketplace.[378]

Advances in the information, financial, and communications technologies have vastly increased the significance of network markets as well as the government's role in regulating these markets. As highlighted in Parts II and III, network technologies are prone to high concentration levels that can enhance consumer welfare through network effects. Therefore, antitrust authorities have had to shift their focus away from market concentration toward anticompetitive tactics such as leveraging market power into new markets and stifling innovation. This Section summarizes the major contours of this shift. Section 1 discusses the evolution of Department of Justice and Federal Trade Commission guidelines for intellectual property licensing. Section 2 discusses significant network market enforcement actions and the challenges of crafting remedies.

1. *Intellectual Property Licensing Guidelines*

Beginning in the 1930's, antitrust regulators took a skeptical view of intellectual property.[379] By the early 1970's, these concerns reached their apex in the U.S. Department of Justice Antitrust Division's "Nine No-No's":

> (1) It is unlawful to require a licensee to purchase unpatented materials from the licensor;
>
> (2) It is unlawful for a patentee to require a licensee to assign to the patentee any patent which may be issued to the licensee after the licensing arrangement is executed;
>
> (3) It is unlawful to attempt to restrict a purchaser of a patented product in the resale of that product;
>
> (4) A patentee may not restrict his licensee's freedom to deal in the products or services not within the scope of the patent;

[378] *See Breakup of the Bell System*, WIKIPEDIA, https://en.wikipedia.org/wiki/Breakup_of_the_Bell_System [https://perma.cc/MS4W-PV4Q].

[379] *See* John DeQ. Briggs, *Intellectual Property and Antitrust: Two Scorpions in a Bottle*, 10 SEDONA CONF. J. 65, 67–68 (2009).

> (5) It is unlawful for a patentee to agree with his licensee that he will not, without the licensee's consent, grant further licenses to any other person;
>
> (6) Mandatory package licensing is an unlawful extension of the patent grant;
>
> (7) It is unlawful for a patentee to insist, as a condition of the license, that his licensee pay royalties in an amount not reasonably related to the licensee's sales of products covered by the patent—for example, royalties on the total sales of products of the general type covered by the licensed patent;
>
> (8) It is unlawful for the owner of a process patent to attempt to place restrictions on his licensee's sales of products made by the use of the patented process; and
>
> (9) It is unlawful for a patentee to require a licensee to adhere to any specified or minimum price with respect to the licensee's sale of the licensed products.[380]

Furthermore, even if a patent-related restraint was not per se unlawful under one of the Nine No-No's, the Department of Justice would still consider bringing an enforcement action if the particular provision was not necessary to the patentee's exploitation of its lawful monopoly and there were less restrictive alternatives to the restrictions that were more likely to foster competition.[381] These enforcement principles focused on attempts by patent holders to extend their patent monopolies to unpatented supplies, to gain control over improvements of their innovations, to determine prices for resale of their patented products, or to engage in market allocations.

With the growing importance of intellectual property assets in the 1970s and 1980s and the dawning of the digital age, economists came to see unconstrained patent licensing as an innovation driver.[382] In 1988, the Antitrust Division shifted from absolute (per se) opposition to licensing restrictions to a "rule of reason" approach to patent licensing that balanced the pro-competitive effects of licensing against potential anticompetitive

[380] *See generally* Bruce B. Wilson, Special Assistant to the Assistant Attorney Gen., Antitrust Division Dep't of Justice, Remarks before the Fourth New England Antitrust Conference: Patent and Know-How License Agreements: Field of Use, Territorial, Price and Quantity Restrictions (Nov. 6, 1970).

[381] *See* Briggs, *supra* note 379, at 65–93; *see also* Wilson, *supra* note 380.

[382] *See* Richard Gilbert & Carl Shapiro, *Antitrust Issues in the Licensing of Intellectual Property: The Nine No-No's Meet the Nineties*, BROOKINGS PAPERS: MICROECONOMICS 283, 286 (1997).

effects in related markets.[383] In 1995, the Department of Justice and the Federal Trade Commission (FTC) expanded upon the 1988 guidelines in crafting the "Antitrust Guidelines for the Licensing of Intellectual Property."[384] These guidelines expressly recognized the generally pro-competitive nature of licensing arrangements, rejected the presumption that intellectual property necessarily creates market power in the antitrust context, and endorsed applying the same general antitrust approach to the analysis of conduct involving intellectual property that the agencies apply to conduct involving other forms of tangible or intangible property.[385]

With the growing role of patents in network industries, the Department of Justice and FTC increasingly recognized the importance of licensing SEPs on fair, reasonable, and non-discriminatory terms.[386] Accordingly, the Department of Justice and the FTC have been far more receptive to patent pools.[387] As Gilbert explains:

> Competition policy toward patent pools has focused on the prevention of anticompetitive practices by patent pool members—individually or collectively through the licensing policies of the pool—and has generally paid little attention to the question of how to encourage the formation and stability of patent pools that benefit consumers. While patent pools have substantial procompetitive benefits when the manufacture or use of products may infringe multiple patents, powerful economic forces prevent beneficial patent

[383] U.S. DEP'T OF JUSTICE, ANTITRUST ENFORCEMENT GUIDELINES FOR INTERNATIONAL OPERATIONS (1988).

[384] *See* U.S. DEP'T OF JUSTICE & FED. TRADE COMM'N, ANTITRUST ENFORCEMENT GUIDELINES FOR INTERNATIONAL OPERATIONS (1995); *see* Letter from Thomas O. Barnett, Assistant Attorney Gen., U.S. Dep't of Justice Antitrust Division, to Michael A. Lindsay, Dorsey & Whitney LLP (Apr. 30, 2007) (on file with the U.S. Dep't. of Justice) [hereinafter Letter from Thomas O. Barnett to Michael A. Lindsay].

[385] *See id.*

[386] *See* U.S. DEP'T OF JUSTICE & USPTO, POLICY STATEMENT ON REMEDIES FOR STANDARDS-ESSENTIAL PATENTS SUBJECT TO VOLUNTARY F/RAND COMMITMENTS (2013); *see generally* Third Party Statement on the Public Interest, Inv. No. 337-TA-745 (June 6, 2012); FTC REPORT, *supra* note 38; Letter from Thomas O. Barnett to Michael A. Lindsay, *supra* note 384.

[387] *See, e.g.*, U.S. DEP'T OF JUSTICE, BUSINESS REVIEW LETTER RELATING TO VMEBUS INTERNATIONAL TRADE ASSOCIATION (2006); Letter from Thomas O. Barnett to Michael A. Lindsay, *supra* note 384; Letter relating to IEEE from Renata B. Hesse, Acting Assistant Attorney Gen., U.S. Dep't of Justice Antitrust Division, to Michael A. Lindsay, Dorsey & Whitney LLP (Feb. 2, 2015); Negotiated Data Sols. LLC (N-Data), F.T.C. File No. 051-0094, 2008 WL 4407246 (Sept. 22, 2008); *In re* Motorola Mobility LLC, F.T.C. File No. C-4410 (2013).

> pools from forming or limit the patents in the pool to only a fraction of the patents that cover the products.
>
> Competition policy should recognize the fragility of patent pools and ensure that patent pool members acting collectively have the same latitude to determine royalties and licensing terms as a single licensor, provided that the pool does not harm lawful competition that would have occurred in the absence of the pool's licenses. In determining which types of patents should be allowed in a pool, competition policy should recognize that a patent pool confers potential benefits if it includes two or more valid complementary patents, and need not harm competition if it has at least one valid patent that is essential to make, sell, or use a product. Inclusion of inessential patents raises potential concerns about foreclosure of alternative technologies and higher royalties for some licenses than would have occurred if these patents were excluded from the pool. However, these concerns should be balanced against the costs of excluding potentially essential patents from the pool.[388]

Disappointingly, the U.S. Dep't of Justice and FTC's updated intellectual property guidelines[389] omit mention of SEPs and FRAND.[390]

2. *Significant Network Market Enforcement Actions*

Notwithstanding the Department of Justice's and FTC's loosening of licensing restrictions, antitrust authorities have pursued several notable enforcement actions in network industries over the past two decades.

In 1996, the FTC alleged that Dell Computer Corporation had violated the Federal Trade Commission Act by failing to disclose its patent rights during the Video Electronics Standards Association standard-setting process and then threatening to enforce those rights against others involved in that process.[391] The resulting consent decree barred Dell from enforcing its patent against computer manufacturers incorporating the pertinent standard.

In 1998, the FTC issued a complaint against Intel Corporation alleging that Intel had sought to maintain its dominance in the microprocessor marketplace by denying essential technical information and product samples of new

[388] Richard J. Gilbert, *Ties that Bind: Policies to Promote (Good) Patent Pools*, 77 ANTITRUST L.J. 1, 3–4 (2010).

[389] *See generally* U.S. DEP'T OF JUSTICE & FED. TRADE COMM'N, ANTITRUST GUIDELINES FOR THE LICENSING OF INTELLECTUAL PROPERTY (2017) (largely reaffirming and modestly updating the 1995 guidelines).

[390] *See* Comments of Law and Business Scholars Submitted to the U.S. Department of Justice and Federal Trade Commission Regarding a Proposed Update to the Antitrust Guidelines for the Licensing of Intellectual Property (Sept. 25, 2016).

[391] *See* Dell Computer Corp., 121 F.T.C. 616 (1996).

microprocessors to companies that, because of intellectual property disputes, had initiated or threatened to initiate litigation against Intel or Intel's customers.[392] The resulting consent decree recognized Intel's right to withhold licenses of its product or information, but limited Intel's ability to retract licenses when customers sought to vindicate its intellectual property rights.

Most significantly, the Department of Justice and eighteen states brought antitrust actions against Microsoft in 1998 alleging that Microsoft's bundling of its browser (Internet Explorer) with its Windows operating system along with restrictive licensing agreements with original equipment manufacturers (such as pricing use of its operating system based on a per processor basis).[393] The government specifically targeted Microsoft's efforts to exclude Netscape from the browser market and to suppress Sun's Java web programming platform. The federal government settled its claims with Microsoft in 2001. [394] The consent decree required Microsoft to share its application programming interfaces with third-party companies and established a process for supervising compliance with the agreement over a five-year period. Nine states proceeded to trial and ultimately implemented somewhat greater oversight over Microsoft's activities.

Crafting a remedy proved especially difficult due to the strong consumer benefits attributable to Microsoft's widely adopted and highly integrated computing platform.[395] Breaking up the company would certainly have caused substantial consumer harm. In the end, the rapid emergence of the Internet Age and mobile computing—along with the ascendance of a new set of competitors such as Google and Facebook, as well as the resurgence of Apple—eroded Microsoft's dominance.

In 2012, the FTC required that Robert Bosch GmbH sell a SPX Service Solutions, a business that makes equipment used to recharge vehicle air conditioning systems, grant licenses to key patents needed to compete in the market for such equipment on the ground that SPX harmed competition by reneging on a commitment to license standard-essential patents on FRAND

[392] *See* Intel Corp., No. 9288, 1999 F.T.C. LEXIS 145 (1999); *see generally* Pitofsky, *supra* note 318.

[393] *See* United States v. Microsoft Corp., No. 98-1232 (D.D.C. May 18, 1998); New York v. Microsoft Corp., No. 98-1233 (D.D.C. May 18, 1998).

[394] *See* United States v. Microsoft Corp., WIKIPEDIA, https://en.wikipedia.org/wiki/United_States_v._Microsoft_Corp.#Settlement [https://perma.cc/4CUB-5FXP].

[395] *See* United States v. Microsoft Corp., 253 F.3d 34, 102 (D.C. Cir. 2001) (reversing the district court order that would have broken Microsoft up because it failed to address Microsoft's contention that such an order would "lower [] rates of innovation and disrupt[] the evolution of Windows as a software development platform").

terms.[396] The FTC declared that "[p]atent holders that seek injunctive relief against willing licensees of their FRAND-encumbered SEPs should understand that in appropriate cases the Commission can and will challenge this conduct as an unfair method of competition under Section 5 of the FTC Act."[397] The Department of Justice has conditioned its approval of acquisitions of substantial patent portfolios by firms with substantial market presence on the commitments to license standard-essential patents on FRAND terms. Prominent examples include: (1) Google's acquisition of Motorola Mobility's portfolio of 17,000 patents and 6,800 patent applications; (2) Apple's acquisition of the nearly 900 patents originally held by Novell and purchased in 2010 by a coalition including Apple, EMC, Microsoft, and Oracle; and (3) acquisition by the "Rockstar" group (made up of Apple, Microsoft, and RIM) of the 6,000 patents and applications made available in the Nortel bankruptcy auction.[398]

VII. ASSESSMENT OF INTELLECTUAL PROPERTY PROTECTION AND COMPETITION POLICY FOR NETWORK TECHNOLOGIES

Drawing on the evolution of intellectual property protection and competition policy explored in Parts V and VI, this Part assesses how the various legal, market, and policy institutions have adapted to the emergence of network technologies in the Information Age. Section VII.A discusses institutional and political economy considerations. Section VII.B then assesses the performance of legal and policy institutions based on the normative principles set forth in Part IV.

A. INSTITUTIONAL CONSIDERATIONS

Intellectual property is not a single, monolithic protective system but rather a complex, overlapping set of protections. Inventors and platform developers can utilize various modes of protecting their innovative endeavors. In addition, they can coordinate with other entrepreneurs to promote and leverage network effects, subject to antitrust constraints.

Protectionist entrepreneurs will naturally exploit the weakest link within the intellectual property chain to gain market advantage. As a result, the

[396] *See* Press Release, Fed. Trade Comm'n, FTC Order Restores Competition in U.S. Market for Equipment Used to Recharge Vehicle Air Conditioning Systems (Nov. 26, 2012).

[397] *See id.*

[398] *See* Michael A. Carrier, *What You Need to Know About Standard Essential Patents*, 8 COMPETITION POL'Y INT'L ANTITRUST CHRON. 6 (2014).

efficacy of the intellectual property system depends critically upon intellectual property gatekeepers—judges, patent examiners, and antitrust enforcers—to ensure that the system coheres.

Therefore, the intellectual property system can be strained and fail to promote balanced protection where critical gatekeepers lack adequate understanding of the overall system or the technologies and economics at issue. The structure of the federal courts creates two opposing vulnerabilities. On the one hand, the regional circuit courts—which handle most copyright and trademark disputes—lack specialization and technological training. They can struggle to understand the complexities of computer software and other technical subject matter in the network technology fields. On the other hand, the Federal Circuit—which handles all patent appeals and some copyright and trademark appeals—is specialized, which can skew their perspective. As numerous scholars have explored, specialty courts, such as the Federal Circuit, are prone to tunnel vision and political capture which could lead to more protectionist interpretations of intellectual property law.[399]

As the Open Handset Alliance and the open source movement have demonstrated, free market institutions can check overbroad intellectual property protection. Entrepreneurs can contract around intellectual property systems in creative ways.[400]

B. Measuring Progress Based on the Normative Principles

The past half century, spanning the birth and ascendancy of the Information Age, has included dramatic evolution of intellectual property protection for network technologies. The process has not always been smooth, but has generally been inclined toward more efficient and effective rules and institutions. Nonetheless, the complexity of the intellectual property system

[399] *See, e.g.*, Rochelle C. Dreyfuss, *The Federal Circuit: A Case Study in Specialized Courts*, 64 N.Y.U. L. REV. 1, 26 (1989) (noting that specialized courts have improved patent law but have significant procedural defects); John R. Allison & Mark A. Lemley, *Empirical Evidence on the Validity of Litigated Patents*, 26 AIPLA Q.J. 185, 241–42 (1998) (noting that specialized courts have improved patent law but have significant procedural defects); Robert P. Merges, *One Hundred Years of Solicitude: Intellectual Property Law 1900-2000*, 88 CALIF. L. REV. 2187, 2224, 2224–33, 2234–39 (2000) (noting that the federal circuit has been pro-patentee, especially in biotechnology, and that it has opened Congress to increased lobbying); WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW 334–53 (2003) (showing the court has been pro-patent in areas like upholding validity, and that this has resulted in an increase in patents); Richard A. Posner & William M. Landes, *An Empirical Analysis of the Patent Court*, 71 U. CHI. L. REV. 111, 128 (2004) (showing the federal circuit is empirically as pro-patent as imagined).

[400] *See generally* Merges, *supra* note 39 (2004) (providing examples of private IP contracting, such as PPIs and the Creative Commons license).

and the dynamism of network technologies has produced persistent pathologies. Fortunately, the flexibility afforded by free market competition and new technologies (such as cloud-based computing) have been valuable antidotes and alternatives to unwarranted intellectual property protection and the accompanying market power. The evolutionary process continues to unfold, and adherence to the key normative principles will benefit from the lessons of the past and ongoing vigilance.

1. *Parsimony Principle*

The parsimony principle aims to promote realization of network benefits by denying intellectual property protection for functional attributes of network technologies absent significant technological advance. This principle comes into conflict with the motivation of some platform developers to control platform development and profit from network effects. Thus, leading platform technology companies advocate robust intellectual property protection for network features of computer software and other technologies through copyright, trademark, and design patent law. These legal regimes do not require assessment of novelty or nonobviousness. In an effort to garner long-lived copyright protection for interface and other software components, they have characterized software code as "high-tech poetry" and analogized computer programs to epic poems and great literature.[401]

Some general jurisdiction judges, with little technical background, were initially receptive to such arguments. They perceived the textual form of software code as more analogous to more conventional literary works than the gears and levers of machines and were less attuned to the broader intellectual property landscape channeling protection for functional features to the utility patent system. Dicta in *Apple v. Franklin* decision opined that "total compatibility with independently developed application programs . . . is a commercial and competitive objective which does not enter into the somewhat metaphysical issue of whether particular ideas and expressions have merged."[402] In *Whelan Associates, Inc. v. Jaslow Dental Laboratory, Inc.*,[403] the Third Circuit's conflation of merger analysis and the idea-expression dichotomy implicitly allowed copyright protection of procedures, processes, systems, and methods of operation that are expressly excluded under § 102(b).

[401] *See* Anthony L. Clapes, Patrick Lynch & Mark R. Steinberg, *Silicon Epics and Binary Bards: Determining the Proper Scope of Copyright Protection for Computer Programs*, 34 UCLA L. REV. 1493, 1500 (1987); *see also* Anthony L. Clapes, *Confessions of an Amicus Curiae: Technophobia, Law and Creativity in the Digital Arts*, 19 U. DAYTON L. REV. 903, 903–04 (1994).

[402] Apple Comput., Inc. v. Franklin Comput. Corp., 714 F.2d 1240, 1253 (3d Cir. 1983).

[403] 797 F.2d 1222 (3d Cir. 1986).

Fortunately, a series of cases in the early to mid-1990s better appreciated the distinction between functionality and creative expression.[404] As a result, while programming a computer can unquestionably be considered "creative" in a general sense, limiting doctrines ensure that the functional aspects are unprotectable under copyright law. The design of an efficient mechanical machine likewise can be creative, but such devices are not eligible for copyright protection unless the aesthetic features can be separated from the functional attributes under the useful article doctrine.[405] Lines of code are the gears and levers of digital machines. The fact that computer software, like a sculptural work, is eligible for copyright protection does not authorize protection for functional features.[406]

Several major technological advances beginning in the mid-1990s deemphasized the role of copyright protection for computer software. The emergence of the Internet as a low-cost, highly scalable distribution ecosystem in the mid to late 1990s vastly expanded the potential for indirect appropriability (e.g., through keyword advertising) and shifted software developers toward open source development. Advances in mobile, Internet-connected digital devices in the early-2000 period paved the way for using software to promote sales of hardware and vastly expanded software distribution through app stores. The new app economy opened a vast array of non-copyright-based business models, such as new forms of advertising (e.g., Yelp). The emergence of cloud-based computing (Software as a Service) reinvigorated digital rights management. These shifts, in combination with the norms that took hold following the *Lotus v. Borland* litigation, produced a period of relative peace with regard to copyright protection of network features of computer software.[407] The parsimony principle prevailed.

[404] *See* Comput. Assocs. Int'l v. Altai, Inc., 982 F.2d 693, 695 (2d Cir. 1992); Sega Enters. Ltd. v. Accolade, Inc., 977 F.2d 1510, 1519 (9th Cir. 1993); Apple Comput., Inc. v. Microsoft Corp., 799 F. Supp. 1006, 1021 (N.D. Cal. 1992), *aff'd in part, rev'd in part*, 35 F.3d 1435, 1445–48 (9th Cir. 1994); Lotus Dev. Corp. v. Borland Int'l, Inc. 49 F.3d 807, 814–15 (1st Cir. 1995), *aff'd by equally divided Court*, 516 U.S. 233 (1996).

[405] 17 U.S.C. § 101 (2018) (definition of "[p]ictorial, graphic, and sculptural works" excludes functional features).

[406] *Id.* § 102(b).

[407] *See* Brian Profitt, *The Impact of Oracle's Defense of API Copyrights*, ITWORLD (Aug. 23, 2011) http://www.itworld.com/article/2738675/mobile/the-impact-of-oracle-s-defense-of-api-copyrights.html [https://perma.cc/B7QG-NT2P] (observing that "[h]istorically, APIs have been regarded as not falling under copyright—the reasoning being that APIs are not creative implementations but rather statements of fact," but also noting the issue had been clouded by the distinction of "open" and "closed"); *see generally* Menell, *supra* note 70, at 651.

That peace was shattered in 2010 with Oracle's filing of a copyright (and patent) infringement lawsuit against Google alleging that the Android operating system infringed copyright protection for the declarations (function names and definitions) in the Java APIs.[408] Drawing on the strategy of the first wave of API copyright litigation, Oracle analogized the labels and code used in the Java APIs to the chapter titles, character names, and plot elements of Harry Potter novels.[409] Based on a questionable interpretation of Ninth Circuit precedent,[410] the Federal Circuit ruled that the structure, sequence, and organization of the 37 Java APIs were copyrightable and remanded the fair use issue for retrial.[411]

Apple's garnering of design patent protection for the rounded, rectangle shape of its iPhone and iPod devices and visual icons also undercut the parsimony principle.[412] These functional elements garnered substantial protection without any showing that they constituted novel and nonobvious technological advances.

These decisions directly undermine the parsimony principle. As a result of the *Oracle v. Google* decision, the safe harbor of clean-room implementation of functional specifications is no longer safe. The *Oracle v. Google* precedent creates the potential for software developers to assert long-lived copyright protection over interface specifications without meeting a substantial threshold of technological advance.

Thus, the *Oracle v. Google* decision warns innovators to steer clear of proprietary software in developing platforms and extensions. Future developers will be careful to avoid using APIs that are vulnerable to copyright assertion. This will reduce the flexibility to join or interoperate with platforms that are not open, but will encourage greater use of open platforms, collaboration, and ex-ante resolution of legal rights. Thus, even though the parsimony principle has been undermined, the flexibility to work around copyright protection through open source and collaborative solutions limits its adverse effects.

[408] *See* Complaint for Patent and Copyright Infringement, Oracle Am., Inc. v. Google Inc., 872 F. Supp. 2d 974 (N.D. Cal. 2012) (No. C 10-03561 WHA), https://docs.justia.com/cases/federal/district-courts/california/candce/3:2010cv03561/231846/1 [https://perma.cc/QV4W-6KST]; Menell, *supra* note 16, at 375–416.

[409] *See* Opening Brief and Addendum for Plaintiff-Appellant at 12–13, Oracle Am., Inc. v. Google, Inc., 750 F.3d 1339 (Fed. Cir. 2014) (No. 17-1118).

[410] *See* Menell, *supra* note 16, at 386–90.

[411] *Oracle Am., Inc.*, 750 F.3d 1339.

[412] U.S. Design Patent Nos. D618,677, D593,087, and D604,305; *see also* Apple Inc. v. Samsung Elecs. Co., 786 F.3d 983 (Fed. Cir. 2015) (affirming lower court's decision that design patents were valid and had been infringed).

2. *Proportionality Principle*

The proportionality principle is the flip side of the parsimony principle coin. Balanced protection for true technological advances in network technologies might be needed to overcome the excess inertia generated by network bandwagons. Patent law provides protection for novel, non-obvious, and adequately disclosed advances in computer systems, processes, and interface design, and other network technologies. Unlike copyright, trademark, or design patent law, utility patent protection protects the functional aspects for network technologies. In theory, therefore, patent protection can provide meaningful protection for overcoming excess inertia. Its efficacy, however, depends on whether it provides the right balance.

In practice, patent protection for interface design and other network technologies has been decidedly mixed. The standards for patent protection might be too low or too high and the duration of protection might be too short or too long to provide the optimal incentive. Moreover, unlike lock-out code, the scope of patent protection does not necessarily align with network features. Furthermore, the costs of pursuing and enforcing patents can distort incentives.

Patent protection of computer software, a principal source of network effects, has experienced a roller coaster over the past four decades. The PTO resisted patent protection for computer software until the late 1960s and only grudgingly afforded such protection in the 1970s and 1980s.[413] The Supreme Court struggled to resolve the eligibility of patent protection for computer software in the 1970s,[414] but ultimately cautiously held that computer programs were eligible in 1981.[415] Nonetheless, software companies were reluctant to pursue such protection, preferring technical protection measures and copyright protection.[416]

[413] *See* Nelson Moskowitz, *The Metamorphosis of Software-Related Invention Patentability*, 3 COMPUTER/L.J. 273, 281–82, 309–11 (1982).

[414] *See, e.g.*, Gottschalk v. Benson, 409 U.S. 63 (1972); Parker v. Flook, 437 U.S. 584 (1978).

[415] *See* Diamond v. Diehr, 450 U.S. 175 (1981).

[416] *See* Menell, supra note 23, at 1346–47, 1351; MacGrady, *Protection of Computer Software—An Update and Practical Synthesis*, 20 HOUS. L. REV. 1033, 1063–64 (1983); ROBERT GREENE STERNE ET AL., THE 2005 U.S. PATENT LANDSCAPE FOR ELECTRONIC COMPANIES 3 (2005)

> The 1980s saw an amazing business phenomena in the U.S. of creation of many start up electronic companies, some of which broke out of the pack of their competitors to become very large companies in their own right. Notable examples are Apple, Microsoft, Oracle, Cisco, Sun, [and] AOL. . . .

Several factors shifted the software industry toward patent acquisition in the early 1990s. Fading hardware companies turned to patent licensing and enforcement campaigns.[417] In addition, some smaller software companies succeeded in enforcing software patents against larger software companies.[418] These developments prompted software companies to pursue defensive patenting.[419] Furthermore, the Federal Circuit liberalized the standards for protecting computers software,[420] just as the Internet (dot-com) era was taking off. This led to a software patenting gold rush in which start-up companies sought patents as signals for raising venture capital and established companies stockpiled patents for defensive purposes.

As discussed in Section V.D, the bursting of the dot-com bubble in 2000 resulted in many software patents falling into the hands of patent aggregators, such as Intellectual Ventures, which produced an unprecedented wave of costly and disruptive patent assertion activity. The low quality and amorphous scope of many of these patents imposed tremendous costs on the software industry and complicated entry into many network technology markets. In addition, new network technologies, such as smart phones, developed in a patent thicket ecosystem.

The effects of patent aggregation and assertion were somewhat alleviated by standard setting organizations requiring FRAND cross-licensing, the emergence of defensive buying funds, such as RPX and Allied Security Trust, and patent pledges.[421] Moreover, the Supreme Court substantially reduced the

> As upstarts, these companies in general did not embrace patents in the slightest.

cf. RIVETTE & KLINE, *supra* note 272, at 41–42 (suggesting ignorance of patent law and antipathy towards software patents as among the reasons companies did not pursue them).

[417] *See generally* MARSHALL PHELPS & DAVID KLINE, BURNING THE SHIPS: INTELLECTUAL PROPERTY AND THE TRANSFORMATION OF MICROSOFT (2009); *see also* ADAM B. JAFFE & JOSH LERNER, INNOVATION AND ITS DISCONTENTS: HOW OUR BROKEN PATENT SYSTEM IS ENDANGERING INNOVATION AND PROGRESS, AND WHAT TO DO ABOUT IT 14–15 (2004); RIVETTE & KLINE, *supra* note 272, at 125.

[418] *See, e.g.*, Lawrence M. Fisher, *Microsoft Loses Case on Patent*, N.Y. TIMES (Feb. 24, 1994), https://www.nytimes.com/1994/02/24/business/microsoft-loses-case-on-patent.html [https://perma.cc/V76Y-EDA9]; Stac Elec. v. Microsoft Corp., No. 93-0413 (S.D. Cal. 1994), *appeal dismissed per stipulation*, 38 F.3d 1222 (Fed. Cir. 1994).

[419] *See* FTC REPORT, *supra* note 38 (discussing defensive patenting).

[420] *See In re* Alappat, 33 F.3d 1526 (Fed. Cir. 1994); State St. Bank & Tr. Co. v. Signature Fin. Grp., 149 F.3d 1368 (Fed. Cir. 1998) (overruling the "business method" exception).

[421] *See generally* Schultz & Urban, *supra* note 259.

risk of injunctive relief,[422] tightened the nonobviousness standard,[423] promoted clearer patent boundaries,[424] and restricted patent eligibility.[425] Congress passed legislation streamlining administrative patent review.[426]

Nonetheless, patent protection for network technologies has proven to be a complex and costly tool for achieving proportional appropriability for network technology innovations. The system has, however, become more balanced and predictable, with improved screening of patent applications, more timely and cost-effective means for invalidating dubious patents through inter partes review at the Patent Trial and Appeal Board, and improved coordination through standard setting and FRAND licensing.

3. *Deterrence Principle*

The deterrence principle stems from, and interacts with, the proportionality principle. Network effects often lead to high market concentration levels, which bring market power with them. The deterrence principle seeks to stunt abuse of such power while promoting network benefits. One of the main antidotes to market dominance by a single platform sponsor is collaboration through standard-setting organizations and licensing agreements, such as FRAND commitments. While such private solutions can promote innovation and downstream competition, they create the potential for anti-competitive behavior.

The past several decades have witnessed substantial evolution of antitrust doctrines and enforcement policies toward a balanced innovation and competitive ecosystem. Antitrust enforcers have come to appreciate the economic benefits of high concentration in network technology markets while also focusing on abusive practices, such as failure to disclose essential patents to standard setting organizations. Standard setting organizations have developed more sophisticated disclosure requirements. In addition, courts

[422] *See* eBay, Inc. v. MercExchange, LLC, 547 U.S. 388 (2006) (holding that the Federal Circuit's general rule that courts should issue permanent injunctions against patent infringers was invalid).

[423] *See* KSR Int'l Co. v. Teleflex Inc., 550 U.S. 398 (2007) (holding that the Federal Circuit's "teaching, suggestion, or motivation" test for nonobviousness was too rigid).

[424] *See* Nautilus, Inc. v. Biosig Instruments, Inc., 572 U.S. 898 (2014) (holding that the Federal Circuit's "amenable to construction" test was insufficiently precise).

[425] *See* Bilski v. Kappos, 561 U.S. 593, 604 (2010) (holding that the "machine-or-transformation" test is not the sole test of patent eligibility); Mayo Collaborative Servs. v. Prometheus Labs., Inc., 566 U.S. 66, 82–85 (2012) (establishing a two-part test for patent subject matter eligibility); Alice Corp. Pty. Ltd. v. CLS Bank Int'l, 134 S. Ct. 2347, 2355 (2014) (refining *Mayo*'s two-part test for patent subject matter eligibility).

[426] American Invents Act, Pub. L. No. 112-29, 125 Stat. 284 (2011).

have broadened their assessment of antitrust, contract, and patent remedies in view of network effects.

The dynamism of network technologies and markets, however, will continue to challenge enforcers, policymakers, and courts. As reflected in the *Sun v. Microsoft* and *Oracle v. Google* litigation, there is a subtle line between promoting interoperability and encouraging innovative forking of established standards.[427]

VIII. FUTURE RESEARCH DIRECTIONS

Following Moore's and Metcalfe's "Laws," network technologies are growing at exponential rates. Digital technologies increasingly drive economic growth. Due in substantial part to the Internet and advances in digital technology, network effects are rapidly diffusing across the economic landscape. Consequently, the interplay of network technologies and intellectual property will continue to evolve rapidly in the coming years and decades.

The opportunities for further research in this field are nearly limitless. Network effects are increasingly important across a growing swath of industries: consumer and industrial products (Internet of Things), energy (smartgrid, autonomous driving, renewable energy), bioinformatics, machine learning, social media, advertising, content creation, and science (database development). The interactions with the range of economic modes (such as contract, business associations, and multi-sided markets), as well as other areas of law (such as privacy and civil liberties) provide a wealth of important research opportunities to explore.

Perhaps most significantly, social media platforms, such as Facebook, Instagram, YouTube, and Twitter, are increasingly important not only to economic activity but also to social mobilization, electoral processes, and the functioning of democracy. These platforms are driven by network effects but are also notable for their polarizing tendencies.[428] These broader ramifications of network effects are critical to legal, social science, and public policy research and reform.

[427] *Cf.* Joseph Farrell, *Compatibility and Competition Policy*, *in* STANDARDS AND PUBLIC POLICY 372, (Shane Greenstein and Victor Stango eds., 2007).

[428] *See* CASS R. SUNSTEIN, #REPUBLIC: DIVIDED DEMOCRACY IN THE AGE OF SOCIAL MEDIA (2017); URI Y. HACOHEN & PETER S. MENELL, UNJUST ENDORSEMENT: HOW SOCIAL MEDIA CORRUPTS COMMERCE AND DEMOCRACY AND WHAT TO DO ABOUT IT (this work is still in process at the time of publication).