COLUMBIA LAW REVIEW

SYMPOSIUM
Common Law for the Age of AI

Mala Chatterjee
Kate Crawford
Mariano-Florentino Cuéllar
Ashley Deeks
Jeanne C. Fromer
Jeannie Suk Gersen
C. Scott Hemphill
Bert I. Huang
Frank Pasquale
Jason Schultz
Katherine J. Strandburg
Tim Wu

VOL. 119 **NOVEMBER 2019** **NO. 7**

COLUMBIA LAW REVIEW

VOL. 119 **NOVEMBER 2019** **NO. 7**

CONTENTS

KEYNOTE

ESSAYS

COLUMBIA LAW REVIEW

VOL. 119 **NOVEMBER 2019** **NO. 7**

ABSTRACTS

ESSAYS

Sex robots are here. Created specifically to allow individuals to simulate erotic and romantic experiences with a seemingly alive and present human being, sex robots will soon force lawmakers to address the rise of digisexuality and the human–robot relationship. The extent to which intimacy between a human and robot can be regulated depends on how we characterize sex with robots—as a masturbatory act, an intimate relationship, or nonconsensual sexual contact—and whether sexual activity with robots makes us see robots as more human or less human. A robot sex panic may be driven primarily by the idea that robots are servile by nature. Critics argue that an inherently nonreciprocal dynamic between humans and robots will translate into exploitative relationships that may fuel abuse of human partners, or that sex robots may further social isolation and retreat from human intimacy. Conversely, sex robots may function as safe—and otherwise unavailable—sexual and emotional outlets for those who may otherwise harm others. They may even train individuals to be more respectful in human relationships. At this point, we do not know how our relationships with robots will inform our relationships with humans, for better or for worse. This Essay explores the consequences of sex robots on society and argues that questions of how sex robots will improve or worsen humans' treatment of one another is the key to regulation to come. What is clear is that sex robots will require us to grapple with our vulnerabilities in relationships, reconsider fundamental rights, and question what it means to be intimate and to be human.

How will we assess the morality of decisions made by artificial intelligence—and will our judgments be swayed by what the law says? Focusing on a moral dilemma in which a driverless car chooses to sacrifice its passenger to save more people, this study offers evidence that our moral intuitions can *be influenced by the presence of the law.*

A recurrent concern about machine learning algorithms is that they operate as "black boxes," making it difficult to identify how and why the algorithms reach particular decisions, recommendations, or

predictions. Yet judges are confronting machine learning algorithms with increasing frequency, including in criminal, administrative, and civil cases. This Essay argues that judges should demand explanations for these algorithmic outcomes. One way to address the "black box" problem is to design systems that explain how the algorithms reach their conclusions or predictions. If and as judges demand these explanations, they will play a seminal role in shaping the nature and form of "explainable AI" (xAI). Using the tools of the common law, courts can develop what xAI should mean in different legal contexts. There are advantages to having courts to play this role: Judicial reasoning that builds from the bottom up, using case-by-case consideration of the facts to produce nuanced decisions, is a pragmatic way to develop rules for xAI. Further, courts are likely to stimulate the production of different forms of xAI that are responsive to distinct legal settings and audiences. More generally, we should favor the greater involvement of public actors in shaping xAI, which to date has largely been left in private hands.

Complex machine learning models derived from personal data are increasingly used in making decisions important to peoples' lives. These automated decision tools are controversial, in part because their operation is difficult for humans to grasp or explain. While scholars and policymakers have begun grappling with these explainability concerns, the debate has focused on explanations to decision subjects. This Essay argues that explainability has equally important normative and practical ramifications for decision-system design. Automated decision tools are particularly attractive when decision-making responsibility is delegated *and* distributed *across multiple actors to handle large numbers of cases. Such decision systems depend on explanatory flows among those responsible for setting goals, developing decision criteria, and applying those criteria to particular cases. Inscrutable automated decision tools can disrupt all of these flows.*

This Essay focuses on explanation's role in decision-criteria development, which it analogizes to rulemaking. It analyzes whether, and how, decision tool inscrutability undermines the traditional functions of explanation in rulemaking. It concludes that providing information about the many aspects of decision tool design, function, and use that can be explained can perform many of those traditional functions. Nonetheless, the technical inscrutability of machine learning models has significant ramifications for some decision contexts. Decision tool inscrutability makes it harder, for example, to assess whether decision criteria will generalize to unusual cases or new situations and heightens communication and coordination barriers between data scientists and subject matter experts. The Essay concludes with

upon to develop common law in response to legal disputes arising out of the deployment of AI.

Many legal scholars have explored how courts can apply legal doctrines, such as procedural due process and equal protection, directly to government actors when those actors deploy artificial intelligence (AI) systems. But very little attention has been given to how courts should hold private vendors of these technologies accountable when the government uses their AI tools in ways that violate the law. This is a concerning gap, given that governments are turning to third-party vendors with increasing frequency to provide the algorithmic architectures for public services, including welfare benefits and criminal risk assessments. As such, when challenged, many state governments have disclaimed any knowledge or ability to understand, explain, or remedy problems created by AI systems that they have procured from third parties. The general position has been "we cannot be responsible for something we don't understand." This means that algorithmic systems are contributing to the process of government decisionmaking without any mechanisms of accountability or liability. They fall within an accountability gap.

In response, we argue that courts should adopt a version of the state action doctrine to apply to vendors who supply AI systems for government decisionmaking. Analyzing the state action doctrine's public function, compulsion, and joint participation tests, we argue that—much like other private actors who perform traditional core government functions at the behest of the state—developers of AI systems that directly influence government decisions should be found to be state actors for purposes of constitutional liability. This is a necessary step, we suggest, to bridge the current AI accountability gap.

Recent advances in machine learning have reinforced the competitive position of leading online platforms. This Essay identifies two important sources of platform rivalry and proposes ways to maximize their competitive potential under existing antitrust law. A nascent competitor is a threatening new entrant that, in time, might become a full-fledged platform rival. A platform's acquisition of a nascent competitor should be prohibited as an unlawful acquisition or maintenance of monopoly. A disruptive incumbent is an established firm—often another platform—that introduces fresh competition in an adjacent market. Antitrust enforcers should take a more cautious approach, on the margin, when evaluating actions taken by a disruptive incumbent to compete with an entrenched platform.

Software has partially or fully displaced many former human activities, such as catching speeders or flying airplanes, and proven itself able to surpass humans in certain contests, like Chess and Go. What are the prospects for the displacement of human courts as the centerpiece of legal decisionmaking? Based on the case study of hate speech control on major tech platforms, particularly on Twitter and Facebook, this Essay suggests displacement of human courts remains a distant prospect, but suggests that hybrid machine–human systems are the predictable future of legal adjudication, and that there lies some hope in that combination, if done well.

Columbia University

SCHOOL OF LAW

LEE C. BOLLINGER, J.D.	*President of the University*
JOHN H. COATSWORTH, PH.D.	*Provost of the University*
GILLIAN LESTER, J.S.D.	*Dean of the Faculty of Law*

THE FACULTY OF LAW

MARK BARENBERG, J.D., M.SC., *Isidor and Seville Sulzbacher Professor of Law; Director, Program on Labor Law and Political Economy*

GEORGE A. BERMANN, J.D., LL.M., *Walter Gellhorn Professor of Law; Jean Monnet Professor of European Union Law; Director, Center for International Commercial and Investment Arbitration*

VINCENT BLASI, J.D., *Corliss Lamont Professor of Civil Liberties*

PHILIP C. BOBBITT, J.D., PH.D., *Herbert Wechsler Professor of Federal Jurisprudence; Director, Center for National Security*

LEE C. BOLLINGER, J.D., *Professor of Law; Seth Low Professor of the University; President of the University*

ANU BRADFORD, LL.M., S.J.D., *Henry L. Moses Professor of Law and International Organization; Director, The European Legal Studies Center*

RICHARD BRIFFAULT, J.D., *Joseph P. Chamberlain Professor of Legislation; Director, Legislative Drafting Research Fund*

JESSICA BULMAN-POZEN, J.D., M.PHIL., *Professor of Law*

ALEXANDRA CARTER, J.D., *Clinical Professor of Law*

SARAH H. CLEVELAND, J.D., M.ST., *Louis Henkin Professor of Human and Constitutional Rights; Faculty Co-Director, Human Rights Institute*

JOHN C. COFFEE, JR., LL.B., LL.M., *Adolf A. Berle Professor of Law*

KIMBERLÉ WILLIAMS CRENSHAW, J.D., LL.M., *Isidor and Seville Sulzbacher Professor of Law; Director, Center for Intersectionality and Social Policy Studies*

LORI FISLER DAMROSCH, J.D., *Hamilton Fish Professor of International Law and Diplomacy*

GIUSEPPE DARI-MATTIACCI, J.S.D., LL.M., PH.D., D.JUR., *Alfred W. Bressler Professor of Law*

BRETT DIGNAM, J.D., *Clinical Professor of Law*

MICHAEL W. DOYLE, PH.D., *University Professor*

ELIZABETH F. EMENS, J.D., PH.D., *Isidor and Seville Sulzbacher Professor of Law*

JEFFREY A. FAGAN, PH.D, *Isidor and Seville Sulzbacher Professor of Law; Professor of Epidemiology*

GEORGE P. FLETCHER, J.D., *Cardozo Professor of Jurisprudence*

MERRITT B. FOX, J.D., PH.D., *Michael E. Patterson Professor of Law; NASDAQ Professor for Law and Economics of Capital Markets; Co-Director, Program in Law and Economics of Capital Markets*

KATHERINE M. FRANKE, J.D., LL.M., J.S.D., *Sulzbacher Professor of Law, Gender, and Sexuality Studies; Faculty Director, Center for Gender and Sexuality Law; Faculty Director, Public Rights/Private Conscience Project*

KELLEN R. FUNK, J.D., PH.D., M.A., *Associate Professor of Law*

PHILIP M. GENTY, J.D., *Everett B. Birch Innovative Teaching Clinical Professor in Professional Responsibility*

MICHAEL B. GERRARD, J.D., *Andrew Sabin Professor of Professional Practice; Director, Sabin Center for Climate Change Law*

RONALD J. GILSON, J.D., *Marc and Eva Stern Professor of Law and Business*

JANE C. GINSBURG, J.D., D.E.A., LL.D., M.A., *Morton L. Janklow Professor of Literary and Artistic Property Law; Director, Kernochan Center for Law, Media and the Arts*

MAEVE GLASS, J.D., PH.D., *Associate Professor of Law*

SUZANNE B. GOLDBERG, J.D., *Executive Vice President for University Life; Herbert and Doris Wechsler Clinical Professor of Law; Co-Director, Center for Gender and Sexuality Law; Director, Sexuality and Gender Law Clinic*

JEFFREY N. GORDON, J.D., *Richard Paul Richman Professor of Law; Co-Director, Millstein Center for Global Markets and Corporate Ownership; Co-Director, Richman Center for Business, Law and Public Policy*

ZOHAR GOSHEN, LL.B., LL.M., S.J.D., *Jerome L. Greene Professor of Transactional Law; Director, Center for Israeli Legal Studies*

MICHAEL J. GRAETZ, LL.B., *Columbia Alumni Professor of Tax Law*

R. KENT GREENAWALT, LL.B., *University Professor*

JAMAL GREENE, J.D., *Dwight Professor of Law*

PHILIP HAMBURGER, J.D., *Maurice and Hilda Friedman Professor of Law*

BERNARD E. HARCOURT, J.D., PH.D., *Isidor and Seville Sulzbacher Professor of Law; Director, Columbia Center for Contemporary Critical Thought; Professor of Political Science*

MICHAEL A. HELLER, J.D., *Lawrence A. Wien Professor of Real Estate Law*

BERT I. HUANG, J.D., *Michael I. Sovern Professor of Law; Vice Dean for Intellectual Life*

CONRAD A. JOHNSON, J.D., *Clinical Professor of Law*

OLATUNDE C.A. JOHNSON, J.D., *Jerome B. Sherman Professor of Law*

KATHRYN JUDGE, J.D., *Professor of Law*

AVERY W. KATZ, J.D., M.A., PH.D., *Milton Handler Professor of Law*

JEREMY KESSLER, PH.D., J.D., M.PHIL., *Professor of Law*

SARAH KNUCKEY, LL.B., LL.M., *Lieff Cabraser Heimann and Bernstein Clinical Professor of Human Rights; Director, Human Rights Clinic; Faculty Co-Director, Human Rights Institute*

JODY KRAUS, J.D., M.A., PH.D., *Patricia D. and R. Paul Yetter Professor of Law and Philosophy; Co-Director, Center for Law and Philosophy*

GILLIAN LESTER, J.S.D., LL.B., *Dean of the Faculty of Law; Lucy G. Moses Professor of Law*

BENJAMIN L. LIEBMAN, J.D., *Robert L. Lieff Professor of Law; Director, Center for Chinese Legal Studies*

JAMES S. LIEBMAN, J.D., *Simon H. Rifkind Professor of Law; Director, Center for Public Research and Leadership*

HON. DEBRA A. LIVINGSTON, J.D., *Paul J. Kellner Professor of Law*

EDWARD LLOYD, J.D., *Evan M. Frankel Clinical Professor in Environmental Law*

CLARISA LONG, J.D., *Max Mendel Shaye Professor of Intellectual Property Law*

HON. GERARD E. LYNCH, J.D., *Paul J. Kellner Professor of Law*

RONALD J. MANN, J.D., *Albert E. Cinelli Enterprise Professor of Law; Co-Director, The Charles E. Gerber Transactional Studies Center*

PETROS C. MAVROIDIS, LL.B., LL.M., DR. JUR., *Edwin B. Parker Professor of Foreign and Comparative Law*

JUSTIN MCCRARY, PH.D., *Paul J. Evanson Professor of Law*

THOMAS W. MERRILL, J.D., *Charles Evans Hughes Professor of Law*

GILLIAN METZGER, J.D., *Stanley H. Fuld Professor of Law*

JOSHUA MITTS, J.D., PH.D., *Associate Professor of Law*

EBEN MOGLEN, J.D., M.PHIL., PH.D., *Professor of Law*

HENRY PAUL MONAGHAN, LL.B., LL.M., *Harlan Fiske Stone Professor of Constitutional Law*

EDWARD R. MORRISON, J.D., M.A., PH.D., *Charles Evans Gerber Professor of Law*

ELORA MUKHERJEE, J.D., *Jerome L. Greene Clinical Professor of Law*

LYNNISE E. PHILLIPS PANTIN, J.D., *Clinical Professor of Law*

KATHARINA PISTOR, LL.M., M.P.A., DR. JUR., *Edwin B. Parker Professor of Comparative Law; Director, Center on Global Legal Transformation*

CHRISTINA D. PONSA-KRAUS, J.D., M.PHIL., PH.D., *George Welwood Murray Professor of Legal History*

DAVID POZEN, J.D., M.SC., *Professor of Law*

JEDEDIAH S. PURDY, J.D., *William S. Beinecke Professor of Law*

ANDRZEJ RAPACZYNSKI, J.D., PH.D., *Daniel G. Ross Professor of Law; Joseph Solomon Professor of Wills, Trusts and Estate Planning*

ALEX RASKOLNIKOV, J.D., M.S., *Wilbur H. Friedman Professor of Tax Law; Co-Chair, The Charles E. Gerber Transactional Studies Center*

JOSEPH RAZ, MAGISTER JURIS, D.PHIL., *Thomas M. Macioce Professor of Law*

DANIEL C. RICHMAN, J.D., *Paul J. Kellner Professor of Law*

CHARLES F. SABEL, PH.D., *Maurice T. Moore Professor of Law*

CAROL SANGER, J.D., *Barbara Aronstein Black Professor of Law*

BARBARA A. SCHATZ, J.D., *Clinical Professor of Law*

DAVID M. SCHIZER, J.D., *Dean Emeritus; Harvey R. Miller Professor of Law and Economics*

ELIZABETH S. SCOTT, J.D., *Harold R. Medina Professor of Law; Vice Dean for Curriculum*

ROBERT E. SCOTT, J.D., S.J.D., *Alfred McCormack Professor of Law; Director, Center on Contract and Economic Organization*

COLLEEN FLYNN SHANAHAN, LL.M., J.D., *Associate Clinical Professor of Law*

MICHAEL I. SOVERN, LL.B., LL.D., D.PHIL., *Chancellor Kent Professor of Law; President Emeritus of the University*

JANE M. SPINAK, J.D., *Edward Ross Aranow Clinical Professor of Law*

SUSAN P. STURM, J.D., *George M. Jaffin Professor of Law and Social Responsibility; Director, Center for Institutional and Social Change*

ERIC TALLEY, J.D., PH.D., *Isidor and Seville Sulzbacher Professor of Law; Co-Director, Millstein Center for Global Markets and Corporate Ownership*

KENDALL THOMAS, J.D., *Nash Professor of Law; Director, Center for the Study of Law and Culture*

KRISTEN UNDERHILL, J.D., D.PHIL., M.SC., *Associate Professor of Law*

MATTHEW WAXMAN, J.D., *Liviu Librescu Professor of Law; Faculty Chair, Roger Hertog Program in Law and National Security*

TIMOTHY WU, J.D., *Julius Silver Professor of Law, Science and Technology*

MARY MARSH ZULACK, J.D., *Clinical Professor of Law*

PROFESSORS EMERITI

VIVIAN O. BERGER, J.D.
BARBARA ARONSTEIN BLACK, LL.B., PH.D.
HARLAN M. BLAKE, J.D., M.A.
HAROLD S.H. EDGAR, LL.B
R. RANDLE EDWARDS, J.D., M.PHIL., A.M.

WILLIAM H. SIMON, J.D.
RICHARD BERENSON STONE, LL.B.
PETER L. STRAUSS, LL.B.
PATRICIA J. WILLIAMS, J.D.
WILLIAM F. YOUNG, LL.B.

AFFILIATED COLUMBIA UNIVERSITY FACULTY

Paul Appelbaum, M.D., *Elizabeth K. Dollard Professor of Psychiatry, Medicine and Law; Director, Division of Psychiatry, Law and Ethics, Department of Psychiatry*

Steven Bellovin, Ph.D., M.S., *Percy K. and Vida L.W. Hudson Professor, Department of Computer Science*

Jagdish N. Bhagwati, Ph.D., *University Professor, Economics and Political Science*

Patrick Bolton, Ph.D., *Barbara and David Zalaznick Professor of Business, Columbia Business School*

Jonathan R. Cole, Ph.D., *John Mitchell Mason Professor of the University; Provost and Dean of Faculties Emeritus*

Lawrence Glosten, Ph.D., M.S., *S. Sloan Colt Professor of Banking and International Finance, Columbia Business School; Co-Director, Program in the Law and Economics of Capital Markets*

Turkuler Isiksel, Ph.D., *James P. Shenton Assistant Professor of the Core Curriculum, Columbia University*

Merit E. Janow, J.D., *Dean, School of International and Public Affairs; Professor of Professional Practice, International Economic Law and International Affairs*

W. Bentley MacLeod, Ph.D., *Sami Mnaymneh Professor of Economics and International and Public Affairs*

Bhaven Sampat, Ph.D., *Associate Professor of Health Policy and Management, Mailman School of Public Health*

VISITORS

Shyamkrishna Balganesh, *Samuel Rubin Visiting Professor of Law*

Emily Benfer, *Visiting Associate Clinical Professor of Law*

Amal Clooney, *Visiting Professor of Law*

Ryan Doerfler, *Nathaniel Fensterstock Visiting Professor of Law*

Matthew C. Jennejohn, *Justin W. D'Atri Visiting Professor of Law, Business, and Society*

Andrew Kent, *Visiting Professor of Law*

Madhav Khosla, *Dr. B.R. Ambedkar Visiting Associate Professor of Indian Constitutional Law*

Catherine YonSoo Kim, *Visiting Professor of Law*

Mariana Pargendler, *Stephen and Barbara Friedman Visiting Professor of Law*

Daniel Raff, *Visiting Associate Professor of Law*

Jessica Roth, *Visiting Professor of Law*

Doron Teichman, *Visiting Associate Professor of Law*

OTHER OFFICERS OF INSTRUCTION

Esinam Agbemenu, J.D.

Kayum Ahmed, M.A., LL.B., LL.M.

Benjamin Alden, J.D.

Akhil K. Amar, J.D.

Susan Amron, J.D.

Kira Antell, J.D.

Keri L. Arnold, J.D.

Barbara Arnwine, J.D.

Eric Askanase, J.D.

Kimberly Austin, PhD

Marion Bachrach, M.A., J.D.

Todd H. Baker, J.D.

Robert Balin, J.D.

Celia Goldwag Barenholtz, J.D.

Rachel Barnett, J.D.

Norman J. Bartczak, Ph.D.

Daniel Berger, J.D.

Saul J. Berman, J.D.

Doris Bernhardt, J.D.

Elizabeth Bernhardt, M.A., J.D., Ph.D.

Sophia Bernhardt, J.D.

June Besek, J.D.

Alison Arden Besunder, J.D.

Bryan Bloom, J.D.

Douglas M. Bregman, J.D.

Katherine Buckel, J.D.

Susanna Buergel, J.D.

Michael Burger, J.D., M.F.A.

Jonathan Bush, J.D.

Toby Butterfield, LL.M., LL.M.

Elizabeth Cabraser, J.D.

Karen Cacace, J.D.

Jennifer Camillo, J.D.

Daniel Capra, J.D.

John Carlin, J.D.,Ph.D.

Steven Chaikelson, M.F.A., J.D.

Wayne Change, J.D.

Hannah Chanoine, J.D.

Steven Charney, M.S., J.D.

Elizabeth Chu, MA., Ph.D.

Christopher Cogburn, J.D.

Jay Cohen, J.D.

Jordana Confino, J.D.

Jennifer Conn, J.D.

Peggy Cross-Goldenberg, J.D.

Shawn Crowley, J.D.

Shannon Cumberbatch, J.D.

Rick D'Avino, J.D.

Jennifer Danis, J.D.

Lev Dassin, J.D.

Anthony E. Davis, M.A., LL.M.

Frederick Davis, J.D.

Owen Davis, M.B.A.

Caroline DeCell, J.D.

Joseph Demarco, J.D.

Timothy DeMasi, J.D., LL.M.

Richard Dicker, J.D., LL.M.

Delyan Dimitrov, J.D.

Joel Dodge, J.D.

Brian Donnelly, M.S., J.D.

Elyse Dreyer, J.D.

Kabir Duggal, LL.M.

Martin Edel, J.D.

Meyer Eisenberg, LL.M.

Anthony Ewing, J.D.

Leslie Fagen, J.D.

Morenike Fajana, J.D.

Kai Falkenberg, J.D.

Louise Firestone, J.D.

Maria Foscarinis, M.A., J.D.

Mavis Fowler-Williams, J.D.

Michael Fox, J.D.

Kevin B. Frankel, J.D.

Andrew Friedman, J.D.

Edward Frischling, J.D.

Hon. Nicholas Garaufis, J.D.

Alejandro M. Garro, J.S.D., LL.M.

Lee Gelernt, J.D.

Michael A. Gerber, J.D.

Michael Gerber, J.D.

Hon. Robert Gerber, J.D.

Hon. Martin Glenn, J.D.

Phillip Goff, M.A., Ph.D.

Martin Gold, J.D., M.P.A.

Janlori Goldman, M.F.A., J.D.

Richard Gray, J.D.

Michelle Greenberg-Kobrin, J.D.

Edward F. Greene, LL.M.

Robert Grey, J.D., LL.M.

Peter Grossi, M.A., J.D.

Y. Shukie Grossman, J.D

John Haggerty, J.D.

Tanya Hajjar, J.D.

Hon. Dorothy Harbeck, J.D.

Robert Harris, M.P.A., J.D.

Christopher Harwood, J.D.

Gail Heatherly, J.D.

Adam Hemlock, J.D.

Patricia Hennessey, J.D.

Jessica Hertz, J.D.

Jay Heubert, J.D., Ph.D.

Jay Hewlin, J.D.

Alexis Johara Hoag, J.D

Eric Hochstadt, J.D.

Silvia Hodges-Silverstein, M.B.A., J.D.

Courtney Hogg, J.D.

Dmitri Holtzman, LL.B.

Kathy Holub, J.D.

Kim Hopper, Ph.D., M.A.

David Paul Horowitz, J.D.

Jeffrey Horowitz, J.D.

Steven Horowitz, M.P.P., J.D.

Corrine Irish, J.D.

Nobuhisa Ishizuka, J.D.

Marc Isserles, J.D.

Alice Izumo, M.S., J.D.

Lawrence Jacobs, J.D.

Jameel Jaffer, J.D.

Joseph Karume Samuel James, J.D.

Dina Jansenson, J.D.

Jessica Jimenez, J.D.

Emily Johnson, J.D.

Larry Johnson, M.P.A., J.D.

O. Thomas Johnson, Jr., J.D.

Bonnie Jonas, J.D.

Patrick Jones, J.D.

Gregory P. Joseph, J.D.

Carl Kaplan, J.D.a

Cathy Kaplan

Jack Kaplan, M.S., M.B.A.

Roberta Kaplan, J.D.

David Kappos, J.D.

Arthur S. Kaufman, J.D.

Risa Kaufman, J.D.

Daniel Kay, J.D.

George Kendall, J.D.

Scott E. Kessler, J.D.

JoAnn Kintz, J.D.

Igor Kirman, J.D.

Jodie Kirshner, J.D.

Edward Klaris, J.D.

Matthew Knecht, J.D.

Saritha Komatireddy, J.D.

Alexandra Korry, J.D.

Joel Kosman, M.S.W, J.D.

Stephen Kotran, J.D.

Susan J. Kraham, J.D, M.U.P.

Hilary Krane, J.D.

Evan Kreiner, J.D.

Alexander Krulic, J.D.

Jo Backer Laird, J.D.

Lisa Landau, J.D.

Ian Laird, LL.B, LL.M.

Gerald Lebovits, J.D., LL.M.

Henry Lebowitz, J.D.

Young Lee, J.D.

Yuanchung Lee, M.A., J.D.

Jay P. Lefkowitz, J.D.

Richard Lehv, J.D.

Dorchen Leidholdt, M.A., J.D.

Jane A. Levine, J.D.

Jennifer Levy, J.D.

Hon. Robert M. Levy, J.D

Allon Lifshitz, J.D.

Lewis J. Liman, M.Sc.,J.D.

Samuel Liss, M.B.A.

Phillipa Loengard, M.S.C., J.D., LL.M.

Walter P. Loughlin, M.A., J.D.

Stephen Louis, J.D.

Christina Ma, J.D.

Jenny Ma, J.D.

Fred Magaziner, J.D.

Gary Mandel, M.B.A., J.D. LL.M.

David Marriott, J.D.

Viren Mascarenhas, J.D., LL.M.

Kathleen Massey, J.D.

Amy McCamphill, J.D.

Amy McFarlane, J.D.

Brendan McGuire, J.D.

James McHugh, J.D.

Ryan McLeod, J.D.

Steven Paul McSloy, J.D.

Gilbert Menna, J.D.

Taryn A. Merkl, J.D.

Janis M. Meyer, M.A..,J.D.

Nicole Mesard, J.D.

James Millstein, M.A., J.D.

Sonya Mirbagheri Cheney, J.D.

Alison Moe, J.D.Tiffany Moller, J.D.

Rahim Moloo, J.D., LL.M

Matthew Morreale, J.D., M.S

Charles Nathan, J.D.

Dana Neacsu, M.L.S., D.E.A., J.D. (Equiv.), L.M.S.

Mariana Newman, J.D.

Mark D. Nielsen, J.D.

Ilan Nissan, J.D.

Michael Nissan, J.D.

Julie North, J.D.

Trevor Norwitz, LL.M.

Delphine Nougayrede, Docteur En Droit

Jenay Nurse, J.D.

Karen Dahlberg O'Connell, J.D.

William Ohlemeyer, J.D.

Victor Olds, J.D.

Eric Pan, M.Sc., J.D.

Michel Paradis, J.D., Ph.D.

Hon. Barrington Parker, LL.B.

Dennis Parker, J.D.

Hillel I. Parness, J.D.

Anar Rathod Patel, J.D.

Rachel Pauley, M.A., J.D.

Susan Paulson, J.D.

Myrna Perez, J.D.

Jill Pilgrim, J.D.

Caroline Polisi, J.D.

COLUMBIA LAW REVIEW

VOL. 119 **NOVEMBER 2019** **NO. 7**

The *Columbia Law Review* convened a symposium in the spring of 2019 to discuss the challenges posed by the extraordinary recent advances in artificial intelligence (AI). The symposium, titled “Common Law for the Age of AI,” brought together a diverse group of scholars to discuss this unique topic. The *Law Review* would like to thank our participants, our faculty sponsors, the Columbia Data Science Institute, and our authors for their fantastic contributions on this important subject.

COLUMBIA LAW REVIEW

VOL. 119 NOVEMBER 2019 NO. 7

KEYNOTE

A COMMON LAW FOR THE AGE OF ARTIFICIAL INTELLIGENCE: INCREMENTAL ADJUDICATION, INSTITUTIONS, AND RELATIONAL NON-ARBITRARINESS

*Mariano-Florentino Cuéllar**

INTRODUCTION

The majority of vehicles on California's vast network of roads make considerable use of information technology.[1] Although most are not yet capable of anything approaching fully autonomous driving, already it is possible to witness something like the following scene. A driver steering one vehicle spies a newer car's reflection in the rear-view mirror. The newer car appears to be driving itself. Whatever the official limits on that sleek vehicle's capability,[2] the person in its driver's seat seems to have no

[*] Justice, Supreme Court of California; Herman Phleger Visiting Professor and former Stanley Morrison Professor, Stanford Law School; affiliated faculty, Center for AI Safety and Freeman Spogli Institute for International Studies, Stanford University. I'm grateful for helpful feedback from Richard Rochman, invaluable conversations with Dario Amodei, Kate Crawford, Geoffrey Irving, Bob Kocher, Fei-Fei Li, and Meredith Whittaker, and superb research assistance from Natalie Heim and Derin McLeod. This Essay is adapted from my keynote address at the *Columbia Law Review*'s 2019 Symposium, "Common Law for the Age of AI."

[1] To some extent, regulators have helped to drive the increasing importance of computing technology in the routine operation of automobiles. See Bill Canis, Cong. Research Serv., R44800, Issues with Federal Motor Vehicle Safety Standards 11–19 (2017), https://fas.org/sgp/crs/misc/R44800.pdf [https://perma.cc/TS69-SHYZ].

[2] See generally David Welch & Elisabeth Behrmann, Who's Winning the Self-Driving Car Race?, Bloomberg (May 7, 2018), https://www.bloomberg.com/news/features/2018-05-07/who-s-winning-the-self-driving-car-race [https://perma.cc/HMC6-8LEN] (noting that the "road to autonomy is long and exceedingly complicated"); Tesla Autopilot—Review Including Full Self-Driving for 2019, AutoPilot Review, https://www.autopilotreview.com/tesla-autopilot-features-review [https://perma.cc/Z5CX-2CPS] (last visited July 29, 2019) (describing Tesla's self-driving capabilities). The extent to which a manufacturer appropriately represents to consumers or regulators the capacity of an autopilot function that falls short of full automation capability can raise plenty of legal issues—under contract law, tort law, and consumer protection statutes and regulations. See, e.g., Edvard Pettersson & Dana Hull, Tesla Sued over Fatal Crash Blamed on Autopilot Malfunction, Bloomberg (May 1, 2019), https://www.bloomberg.com/news/articles/2019-05-01/tesla-sued-over-fatal-crash-blamed-on-autopilot-navigation-error [https://perma.cc/VFX7-94UF]. Indeed, it's far from clear whether a concept such as "full" automation is even viable when functions that humans

interaction with the steering wheel when the driver of the older vehicle begins observing. Instead, the person in the driver's seat of that car is engaged in a mix of what seems like personal grooming, texting, and distracted glancing out the side window. Almost subconsciously, the driver of the older car realizes he is tweaking his own driving to test (within the limits of what's safe, of course) the way the algorithm appears to be driving the car behind him. If the older car slowed down or applied the brakes, the newer car behind would slow—gently if the front car decelerated slowly, and somewhat more suddenly if the driver of the older car applied the brakes more unexpectedly. Then the driver of the older vehicle realizes that if he stops for traffic and waits for the car in front to advance a bit before quickly accelerating, the autopiloted car stays behind and opens up a gap in traffic, tempting drivers in other lanes to switch into the opened-up spot. But if the driver of the older car speeds up more gradually, the newer vehicle stays close to the older car. So the older car's driver could effectively tighten the invisible coupling between his car and the more autonomous one or break it based on the rate of acceleration. Finally, when the lane next to the older car is clear, the driver realizes that a slight deviation in how centered his car is in the original lane achieves something significant—it seems to make the autopilot in the newer car behind disengage, forcing that driver to take over the steering wheel.

Even these few seconds of reciprocal steering and autopiloting on a California freeway tell a story: Simple choices can shape complex norms about how we rely on our machine infrastructure. More than simply emphasizing the importance of intricate algorithmic details affecting vehicular behavior, these stories also underscore how much humans are witnessing the steady integration of manufactured intelligence into everyday social life.[3] No doubt a human driver can feel like the Oscar Isaac character dancing with the robot in the film *Ex Machina*.[4] Sometimes this means that humans will be shaped in subtle but potentially enormously consequential ways by artificial intelligence (AI) techniques affecting the flow of information, the distance between cars, or the timing of persua-

colloquially bundle into a single category, such as driving, are easily disaggregated into distinct sub-functions that may call for different automation processes or degrees of human interaction, and when consumers routinely use available technologies in ways that fail to correspond to prescribed limits.

[3] See Meredith Whittaker et al., AI Now Report 2018, at 10–11 (2018), https://ainowinstitute.org/AI_Now_2018_Report.pdf [https://perma.cc/JL95-7XKH] (describing the variety of settings where people routinely interact with systems displaying characteristics of artificial intelligence, and the broad range of functions performed); Ted Greenwald, What Exactly Is Artificial Intelligence, Anyway?, Wall St. J. (Apr. 30, 2018), https://www.wsj.com/articles/what-exactly-is-artificial-intelligence-anyway-1525053960 (on file with the *Columbia Law Review*) (same).

[4] Ex Machina (Film4 & DNA Films 2014).

sive messages, for example.[5] Yet when we share the road, and indeed the world, with artificially intelligent systems, the direction of influence can also run in the opposite direction: Influencing the performance of an AI system need not be a very elaborate, high barrier-to-entry activity. The aforementioned driver's heavily analog, twentieth-century methods did fine in controlling, to some extent, a complex amalgam of software and hardware that is almost certainly also susceptible to—if surely somewhat tightly secured against—more sophisticated hacking.[6] Indeed, the co-evolution of human and artificial intelligence—what we could call our dance with machines—is well on its way to becoming routine. The dance continues as we navigate artificial chatbots, insurance transactions, court avatars, earnest advertising appeals, and borders.

Lurking in the background is law, along with the assumptions and norms it helps sustain. That this dance is playing out in the world's most economically complex and geopolitically powerful common law jurisdiction—the United States, still the preeminent hub for innovation in AI[7]—makes it appropriate to explore what relevance the common law and AI hold for each other. In fact, even accounts of American law that foreground the administrative state retain a prominent if not starring role for the system of incremental adjudication associated with American common law. Indeed, the roads, buildings, and corners of cyberspace where humans are increasingly interacting with manufactured intelligence also reveal another development of considerable importance for lawyers and judges: AI is becoming an increasingly relevant development for the American system of incremental, common law adjudication. The design of a vehicle with some capacity for autonomous driving can spur contract and tort disputes with qualities both familiar and novel.[8] Even decades

[5] See, e.g., Robert M. Bond et al., A 61-Million-Person Experiment in Social Influence and Political Mobilization, 489 Nature 295 (2012) (finding that randomly assigned political mobilization Facebook messages influenced Facebook users' offline political activity).

[6] Lying somewhere in between sophisticated cybersecurity intrusions and easily deployed human-driven techniques to control AI systems is the use of adversarial attacks to disrupt the expected operations of machine learning systems. See, e.g., Alexey Kurakin, Ian J. Goodfellow & Samy Bengio, Adversarial Machine Learning at Scale 1–2 (2017), https://arxiv.org/pdf/1611.01236.pdf [https://perma.cc/2XBM-UVD2] ("[N]eural networks and many other categories of machine learning models are highly vulnerable to attacks based on small modifications of the input to the model at test time . . . .").

[7] See Sarah O'Meara, China's Ambitious Quest to Lead the World in AI by 2030, 572 Sci. Am. 427, 428 (2019) ("Most of the world's leading AI-enabled semiconductor chips are made by US companies such as Nvidia, Intel, Apple, Google and Advanced Micro Devices.").

[8] See, e.g., Mark A. Geistfeld, A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation, 105 Calif. L. Rev. 1611, 1632–74 (2017) (discussing manufacturer liability for autonomous vehicle crashes and hacks); Bryant Walker Smith, Automated Driving and Product Liability, 2017 Mich. St. L. Rev. 1, 32–56 (discussing products liability and personal injury litigation in the context of autonomous vehicles); see also Jack Boeglin, The Costs of Self-Driving Cars: Reconciling

ago, American courts were sometimes already facing legal questions foreshadowing dilemmas one can reasonably expect the present century to serve up about the balance of human and machine decisionmaking. A court in Arizona, for example, was forced to resolve whether punitive damages could be imposed on a transportation company, which failed to use information technology to track the work of its drivers and limit them from working excessive hours.[9] Just as courts once had to translate common law concepts like chattel trespass to cyberspace,[10] new legal disputes—turning on subtle distinctions revealed by digital evidence of neural-network evolution that bear on a party's responsibility for causing harm, for example—will proliferate as reliance on AI becomes more common. The reasonableness of a driver's decision to rely on a vehicle's autonomous capacity, or an organization's choice to delegate a complicated health or safety question to a neural network, will almost certainly spur a new crop of disputes in American courtrooms.

Given the speed and importance of these developments, my purpose here is to begin surveying the fertile terrain where the American system of common law adjudication intersects with AI. American society depends both on technology and the role of incremental common law adjudication in the legal system. The growing importance of AI gives us reason to consider how AI, common law, and society may affect each other. In particular, such an exploration should take account of the common law's role as a default backstop in social and economic life in the United States and a number of other major economies. Even beyond the strict doctrinal limits of torts, property, and contracts, common law ideas tend to set the terms for conversations among elites and even the larger public about the way social and economic interactions ordinarily occur, and how public agencies should analyze the problems—ranging from financial regulation to occupational safety—they are designed to mitigate.[11] Beyond serving as a default means of structuring interactions and a framework for analyzing social and economic life, the common law also offers an apt metaphor for how law, society, and technological change affect each other over the drawn-out process of applying broad social commitments to specific fact patterns. So it is no surprise that any

Freedom and Privacy with Tort Liability in Autonomous Vehicle Regulation, 17 Yale J.L. & Tech. 171, 174–75, 185–201 (2015) (noting the "uncertainty surrounding the complex liability issues for crashes involving [autonomous vehicles], which, in many ways defy the traditional conceptions of fault and agency at play in automobile accidents").

[9] See Torres v. N. Am. Van Lines, Inc., 658 P.2d 835, 838–39 (Ariz. Ct. App. 1982).

[10] See Intel Corp. v. Hamidi, 71 P.3d 296, 308 (Cal. 2003) (declining to find that emails from a former employee to numerous current employees criticizing the company's employment practices could, despite their unauthorized nature, constitute trespass to chattels).

[11] See, e.g., Mariano-Florentino Cuéllar, Administrative War, 82 Geo. Wash. L. Rev. 1343, 1439 (2014) (discussing how ideological norms and the common law appeared to buttress each other and reinforced concerns about government ownership of industry during wartime mobilization in the early 1940s).

intellectually candid conversation about law and AI—particularly in the United States—must be to a considerable extent a conversation about the relationship between AI and the common law.

After defining some terms and setting the stage, I offer three preliminary ideas. First, our society already regulates AI through a backstop arising from the common law—and rightly so. Second, some degree of explainability that is well-calibrated to foster societal deliberation about consequential decisions is foundational to making any AI involved in human decisionmaking compatible with tort and other common law doctrines. At least one version of this ideal that merits attention could be termed "relational non-arbitrariness" to foreground the importance of buttressing—through both the common law and public law—society's capacity to deliberate about, and revise, the process through which it makes the choices that matter most. Finally, common law doctrines have room to integrate societal considerations involving organizational realities and institutional capacity, and concerns about matters such as the erosion of human knowledge that would be risky to ignore.

I. THE SCOPE OF A SHARED CONVERSATION ABOUT LAW AND AI: POINTS OF DEPARTURE

One can think of society as the aggregate of people who live together in a more or less ordered community. The very idea of society implies at least some degree of fairly constant change. Just as people across generations are defined by their evolving relationships to different groups or formal organizations,[12] those same people and the organizations with which they are affiliated are defining, through their behavior, their bonds with the increasingly adaptive technologies that surround them. For two reasons, close observation of the legal system and its common law component proves a revealing method to discern how some of that change happens. For one, the lawyers, clients, judges, and policymakers working through or within the legal system often play a part—sometimes a pivotal one—in the struggles over how society evolves.[13] But

[12] See Max Weber, The Theory of Social and Economic Organization 118–20 (Talcott Parsons ed., A.M. Henderson & Talcott Parsons trans., 1947) (explaining the shifting nature of various kinds of social relationships); see also Raymond Geuss, History and Illusion in Politics 14–20 (2001) (explaining how a single "political association" can shift over time); Sheldon S. Wolin, Max Weber: Legitimation, Method, and the Politics of Theory, 9 Pol. Theory 401, 409–10 (1981) (explaining that Weber's definition of "culture" was concerned with social "meaning" and "patterns").

[13] For examples of how key actors within the legal system affect policy outcomes through an alchemy of discretionary choices, legal interpretations, and strategies for reform of institutions and their legal authority, see Mariano-Florentino Cuéllar & Keith Humphreys, The Political Economy of the Opioid Epidemic, 37 Yale L. & Pol'y Rev. (forthcoming 2019) (manuscript at 48–53) (on file with the *Columbia Law Review*) (analyzing these dynamics in policymaking and litigation associated with opioid abuse); Mariano-Florentino Cuéllar, Refugee Security and the Organizational Logic of Legal Mandates, 37 Geo. J. Int'l L. 583, 587 (2006) (exploring this process in international law); Mariano-

whether any particular actors playing their part in the legal system fail or succeed at their goals, the legal opinions, legislative enactments, and administrative materials that memorialize much of the legal system's work also tend to reveal a story of society's public narrative of justifications and rationales for action: its compromises and aspirations.[14] That disputes about property interests, contract formation, products liability, and other aspects of the legal system can reveal so much is not only reason to take seriously the ideas and internal dynamics that define its work as a description of societal conflict and change; it's also a reason to retain humility about what deeper normative insights that may be persuasive across different segments of society—or even across cultures—can be gleaned from it. Bearing in mind that spirit of humility, we can at least observe in the American common law tradition and its related statutory or regulatory developments some insights about ideas that may be valuable amidst the transitions currently under way.

The enormous changes in the alchemy of algorithms and data, in social norms about computing, and in the resources available for technological development suggest two scenarios that may arise in the next few decades with respect to those transitions. One scenario takes as its point of departure the still-substantial limitations that bedevil many aspects of AI technology well into the twenty-first century—including in domains such as natural language processing and complex motor functions in robotics. Under this scenario, steady but gradual change occurs in AI as well as in the norms, institutions, and financial arrangements affecting its use. As we further leverage reinforcement learning and its variants, alongside conventional uses of supervised and unsupervised learning, the broad outlines of our world could remain much the same as they are now. Lawyers and their clients will continue navigating familiar disagreements about domestic and international politics. Whether they graduate from college or struggle to even finish high school, young people will navigate a labor market in which unemployment is a problem—but a manageable one in advanced countries—and relationships are primarily among humans. The autonomous vehicles in that scenario change only slowly from the one behind me in the recollection I shared; the rates at which its successors improve are limited by constraints of money, physics, lack of human imagination, and familiar global developments like recessions and climate change. AI is in that picture, but the difference compared to how it works now isn't categorical: It's still mostly a technology

Florentino Cuéllar, The Tenuous Relationship Between the Fight Against Money Laundering and the Disruption of Criminal Finance, 93 J. Crim. L. & Criminology 311, 403–04 (2003) (discussing this process in the criminal and national security context).

[14] See Gerald E. Frug, The Ideology of Bureaucracy in American Law, 97 Harv. L. Rev. 1276, 1279–86 (1984) (analyzing corporate and administrative law "as a series of stories that assure us about the acceptability of bureaucratic organizations"); Duncan Kennedy, The Structure of Blackstone's *Commentaries*, 28 Buff. L. Rev. 205, 210, 214–16 (1979) (describing categories of legal reasoning as "social construction[s]").

of massive data analyzed using some kind of artificial neural network deploying enormous computing, where some tasks like automated translation and clustering get faster and become more ubiquitous, but—to channel Richard Haass—the world of the future is much like the world of the present.[15]

But in another scenario, the next one or two decades are quite discontinuous relative to our present. Here some mix of cheaper and greater computing power, innovation in designing algorithms, and our understanding of intelligence prove far more transformative over the next two decades or so. It becomes possible to imagine a world where some material subset of the population has deep emotional attachments to AI systems; where far more of the language we respond to or learn from is artificially generated; where some forms of friendship and work attachments are commodified through AI; where many major decisions about resources, entertainment, coercion, or innovation are routinely made with almost no human intervention; and where labor markets bear little resemblance to present ones. The difference between these scenarios is not the main subject of my talk, though it lurks in the background.

I suspect the distinction between these two scenarios turns heavily on several almost certainly interrelated questions: two technical and one social. The two technical questions that loom large are whether (1) enough progress occurs in natural language processing to simulate routine human communication of medium-to-high complexity (whether written or spoken); and (2) whether we scale the availability of reliable autonomous transportation. The social question is whether norms about the value of human decisionmaking, and the propriety of quite complex, emotionally meaningful communication and relationships with AI relative to humans, shift in favor of even more robust acceptance of AI-driven decisions and interaction.

That these scenarios are distinct in important ways should not obscure a crucial point of convergence. Whether drastic changes in employment or social norms about our relationship to machines occur in the next two decades or take longer, the legal system in general—and the common law in particular—will be a major focal point for certain pronounced societal dilemmas associated with AI. We can better understand those choices not only by recognizing the common law's role as a regulatory backstop but also by focusing attention on the centrality of reasoned deliberation across people and institutions, at least in the American legal tradition. And because the aspirations associated with the legal system inevitably run the risk of encountering detours and roadblocks, we must also acknowledge how much institutions matter—both as the targets for

[15] See generally Richard N. Haass, Where to Go from Here: Rebooting American Foreign Policy, Foreign Aff., July/Aug. 2017, at 2, 9 (arguing that "the old challenges have not gone away," despite technological advances and increased globalization).

much of legal doctrine and also as the means through which the legal system implements much of its work.[16]

In this context, the term artificial intelligence means information technology that learns in some way, can perform some functions we consider to require intelligence if performed by humans, and provides functions or outputs that non-experts would consider trusting. Framed this way, the term encompasses both the domain-specific applications performing specific functions involving financial analysis or autonomous driving, for example, as well as systems aiming to simulate general intelligence through conversation or analytic capacity across domains. This description pivots, too, on the presence of a distinction between AI and conventional statistical inference—though obviously specific machine learning techniques at the heart of certain AI applications implicate both computer science and statistics.[17]

II. INCREMENTAL COMMON LAW ADJUDICATION AS THE DEFAULT FORM OF SOCIAL REGULATION

In a market economy with our historical tradition, the common law is the default framework for making sense of social and economic interaction.[18] Its conventions have informed, and indeed predated, the rise of the modern administrative state. The common law's influence is therefore powerful not only in its direct consequences for discrete transactions, such as the buying or selling of land, but in the ideas it's buttressed about who owes what to whom and for what reason—what duty of care, for example, two people owe each other, and thus what features of social life call for some judge-made or administrative remedy. Which means that in some sense, the pervasive common law backstop to social life provides a kind of first-draft regulatory framework—however imperfect—for managing new technologies ranging from aviation to email. Despite sometimes strong protestations to the contrary (especially from

[16] See Mariano-Florentino Cuéllar, Beyond Weber: Law and Leadership in an Institutionally Fragile World, 69 Stan. L. Rev. 1781, 1785–87 (2017).

[17] See Michael Jordan, Artificial Intelligence—The Revolution Hasn't Happened Yet, Medium (Apr. 19, 2018), https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7 [https://perma.cc/56CM-MLPR] ("The developments which are now being called 'AI' arose mostly in the engineering fields associated with low-level pattern recognition and movement control, and in the field of statistics—the discipline focused on finding patterns in data and on making well-founded predictions, tests of hypotheses and decisions.").

[18] See Grant Gilmore, The Ages of American Law 3–4 (2d ed. 2014); see also Francesco Parisi, The Efficiency of the Common Law Hypothesis, *in* The Encyclopedia of Public Choice 519, 519 (Charles K. Rowley & Friedrich Schneider eds., 2004); Rafael La Porta, Florencio Lopez-de-Silanes, Andrei Shleifer & Robert W. Vishn, Law and Finance, 106 J. Pol. Econ. 1113, 1115–17 (1998).

my home community in Silicon Valley[19]), society already "regulates" AI. That's true even in the absence of statutes and regulatory rules governing AI (though some of those exist, too, particularly in the autonomous vehicle context[20]). But the ultimate regulatory backstop here is the common law.

For thoughtful lawyers practicing in the American tradition, this may seem like a commonplace or even banal observation. But consider the implications: Through tort, property, contract, and related domains, society will shape how people use AI and will define what it means to abuse AI. We may draw on property law to address whether an annoyingly chatty AI system programmed to seek what it defines as autonomy, or a close affiliation with another person it claims has influenced its reasoning, can be "owned" and to what end. Contract law will be one setting in which we work on resolving whether the perfect, AI-spiked pitch for a bargain tailored and timed to overwhelm someone's judgment is unconscionable.[21] All this will indirectly shape norms and semiformal alternative dispute resolution systems, and it will play out even—indeed, especially—if statutes or regulatory rules meant to comprehensively regulate AI "sectors" remain on the shelf gathering dust. Sometimes the risk of disparate doctrinal decisions across many states or doctrines that cut against economic interests can prompt efforts to preempt the common law. But given the context of a relatively robust American federalism, it's not always easy to preempt state-level common law decisionmaking. Consider, for example, how even sweeping federal autonomous vehicle legislation carves out a robust domain (presumably on "laboratories of democracy" type grounds) for states to make common law decisions through the courts about autonomous vehicles.[22] We may also expect criminal law—that ever demanding stepchild of the common law and our statutory present—to deliver a share of dilemmas with interesting trade-offs.

[19] See, e.g., Tom Simonite, Google Says It Wants Rules for the Use of AI—Kinda, Sorta, WIRED (Feb. 2, 2019), https://www.wired.com/story/google-says-wants-rules-ai-kinda-sorta/ [https://perma.cc/W2KZ-FTHW].

[20] See Jack Karsten and Darrell West, The State of Self-Driving Car Laws Across the U.S., Brookings Inst.: Techtank (May 1, 2018), https://www.brookings.edu/blog/techtank/2018/05/01/the-state-of-self-driving-car-laws-across-the-u-s/ [https://perma.cc/P2TE-F48V].

[21] See, e.g., Maxwell v. Fid. Fin. Servs., Inc., 907 P.2d 51, 53–54 (Ariz. 1995); Perdue v. Crocker Nat'l Bank, 702 P.2d 503, 512–13 (Cal. 1985).

[22] See, e.g., David H. Coburn, Dane Jaques & Anthony J. LaRocca, Senate Commerce, Science, and Transportation Committee's AV START Act Advances, Steptoe (Oct. 11, 2017), https://www.steptoe.com/print/content/20768/Senate-Commerce-Science-and-Transportation-Committees-AV-START-Act-Advances.pdf?q= [https://perma.cc/LU8K-KC3V] (noting that "the AV START Act also preserves the existing rule that compliance with a federal safety standard does not exempt a person from common law liability under state law"); Aarian Marshall, Congress Races to Pass a Self-Driving Car Law by Year's End, WIRED (Dec. 5, 2018), https://www.wired.com/story/av-start-act-senate-congress-new-language-self-driving/ [https://perma.cc/88E9-YWHF] (noting that the bill clarifies "which level of government controls what part of self-driving car testing and operations").

It's worth reflecting on what society's frequent inability to recognize the common law as a form of "regulation" may suggest. Obviously important distinctions lurk in the difference between the common law and statutes or regulatory rules. And as Frank Pasquale's essay for this Symposium reminds us, interaction between regulatory rules and the common law can have a powerful effect on specific substantive domains ranging from bank finance to insurance to family relationships.[23] Yet we may consider what this blind spot about how our legal system operates says about how even informed laypeople ordinarily perceive the market economy: as having a nonregulatory default. This is not only deeply misleading; it misses the status quo and empowers certain institutions—in this case state courts, for example—that may be in some ways less and in other ways more prepared for the challenge than even they realize.

III. GIVING MEANING TO OPEN-ENDED COMMON LAW CONCEPTS BY REDUCING OPACITY: RELATIONAL NON-ARBITRARINESS

Whether the common law exerts an influence on AI primarily through courts or through common law-inspired norms shaping the public's understanding of social and economic relations, we can glean some insight into the evolving place of AI in society by training attention on some of the common law's most prominent doctrinal building blocks. Some of its most crucial components, such as the concept of reasonable care in tort law, presuppose an ability for a decisionmaker (such as a court) to observe sufficient contextual details to gauge objective reasonableness. As the use of AI becomes more common to design content and display information, and to inform, and in some cases essentially make, socially important decisions such as whom to hire and how to interact with the market, the capacity to assess context by understanding how AI systems make decisions in order to determine reasonableness becomes particularly important. Let me illustrate the point by delving more deeply into tort law.

Anglo American tort law relies on concepts of proximate causation, foreseeability, and duty, which together provide a more adaptive alternative to many forms of less flexible regulation meant to force parties to internalize the social costs of their actions and decisions.[24] Because the nature of the proximate causation and often-related foreseeability inquiries are flexible enough to take account of changing social, technological, and economic conditions, resolution of cases in this area often involves an interplay between application of well-settled principles and flexibility for the doctrine to adapt to the concerns raised by new situa-

[23] See Frank Pasquale, Data Informed Duties in AI Development, 119 Colum. L. Rev. 1917, 1928–31 (2019) (describing the interaction between judge-made common law and expert-led agencies).

[24] See, e.g., S. Cal. Gas Leak Cases, 441 P.3d 881, 887–88 (Cal. 2019).

tions. Evaluating these decisions will raise some intricate flexibility–fidelity trade-offs for our profession.

Those trade-offs will arise whether decisions involve routine matters such as driving, or more specialized activity such as detecting suspicious transactions. As a number of you have anticipated in papers and comments, one concern is that evaluating proximate causation requires a certain understanding (and ultimately, explainability) of the rationales for AI-driven decisions, without which it's difficult to complete in any defensible way the conventional doctrinal inquiry (because, at a minimum, it's not clear how justifiable it is for a person to rely on a particular kind of decisionmaking technology). So making concepts like reasonableness relevant in a world more reliant on AI systems will depend to some extent on the intricate design choices affecting how AI systems exchange information with humans, and how decisionmaking "justifications" are somehow extracted from artificial neural networks and similar systems. The challenge here is in some ways not so different from what happens when AI technology informs judgment about open-ended statutes like the Administrative Procedure Act or constitutional concepts like reasonable suspicion.[25] The most likely way to make an infrastructure of machine decisionmaking (or at least decision support) conform to a system with the aspirations of our own is to expect machine answers to conform to what I call "relational non-arbitrariness"—a concept not unrelated to what Ashley Deeks calls xAI,[26] and perhaps an example of Kate Strandburg's point about how human decisionmaking tends to be shared decisionmaking.[27]

Rooted to some extent in familiar concerns about shared deliberation[28] as well as decision costs,[29] relational non-arbitrariness calls for

[25] See, e.g., Cary Coglianese & David Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era, 105 Geo. L.J. 1147, 1170–76 (2017) (discussing "adjudicating by algorithm" and "rulemaking by robot"); Andrew Guthrie Ferguson, Big Data and Predictive Reasonable Suspicion, 163 U. Pa. L. Rev. 327, 351 (2015) (describing how big data will change policing and suspicion of criminal activity).

[26] See Ashley S. Deeks, The Judicial Demand for Explainable Artificial Intelligence, 119 Colum. L. Rev. 1829, 1829–30 (2019); Ashley S. Deeks, Predicting Enemies, 104 Va. L. Rev. 1529, 1569 (2018); see also Matt Turek, Explainable Artificial Intelligence (XAI), Program Information, DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence [https://perma.cc/N23J-KSWS] (discussing explainable AI and explainable machine learning as a concept) (last visited Aug. 26, 2019).

[27] Katherine J. Strandburg, Rulemaking and Inscrutable Automated Decision Tools, 119 Colum. L. Rev. 1851, 1854–55 (2019).

[28] See, e.g., John Dewey, The Public and Its Problems: An Essay in Political Inquiry 174–91 (Melvin L. Rogers ed., Ohio Univ. Press 2016) (1927) ("[Democracy] is the idea of community life itself"); Jurgen Habermas, Between Facts and Norms 126–28 (1996) ("The citizens themselves become those who deliberate and, acting as a constitutional assembly, decide how they must fashion the rights that give the disclosure principle legal shape as a principle of democracy."); Alexander Meiklejohn, Political Freedom: The Constitutional Powers of the People 24–28 (1965) (examining the freedom of speech and discussion in the context of "the traditional American town meeting").

the evaluation of how not only private organizations but public institutions make decisions. It takes seriously that decision costs should not routinely swamp the benefits of policies or legal rules, and depends on first considering whether some basis exists for a decision made by a human in close consultation with an AI system or by the system itself, in principle, such that we can defend the decision as non-arbitrary. Second, it calls for asking whether the relationship between the machine or analytical tool and the human conveys some of the complexity involved in the analysis and competing values at stake in the decision. And third, it calls for asking whether the process for making decisions supports further deliberation about the decision among some members of the community of people involved in or affected by it.

Admittedly, relational non-arbitrariness is a mouthful, but then so are terms like reasonable foreseeability and joint and several liability. The point of using a broader term is to think across fields like torts and civil procedure, regulation and constitutional law—and to ask whether an explanation is sufficient to let lawyers and informed laypeople engage in meaningful conversations about how a decision was made and for what reason. A touch of explanation will help you see how much it's already part of our law's fabric: It's taking seriously tort cases like *Biakanja v. Irving* that call for consideration of factors like prevention of future harm and moral blame to decide on the existence of a duty of care,[30] and seeking clarity about the assumptions underlying the arguments about these factors. It's considering whether the justifications offered by an agency arguing its conduct was supported by substantial evidence generalize to other contexts, or at least disclose the extent of the fit between an agency's argument and the relevant legal standard. It's expecting that a hearing to satisfy due process will involve enough transparency to know whether a decisionmaker has effectively delegated all power to an algorithm.[31]

A focus on reasoned explanation and public deliberation is perhaps especially prominent in public law. One can readily discern a judicial concern not only with reason-giving but also with how reason-giving permits deliberation and decisionmaking about accountability in the Supreme Court's recent majority opinion in *Department of Commerce v.*

[29] See Mariano-Florentino Cuéllar, Auditing Executive Discretion, 82 Notre Dame L. Rev. 227, 251–52, 256–57 (2006).

[30] 320 P.2d 16, 19 (Cal. 1958).

[31] See State v. Loomis, 881 N.W.2d 749, 774–76 (Wis. 2016) (Abrahamson, J., concurring) (arguing that courts using risk assessment tools to inform sentencing decisions "must set forth on the record a meaningful process of reasoning addressing the relevance, strengths, and weaknesses" of the tool); see also Michael T. v. Crouch, No. 2:15-CV-09655, 2018 WL 1513295, at *3, *9–10 (S.D. W. Va. Mar. 26, 2018) (analyzing the use of algorithms to determine individualized budgets for state disability benefit recipients).

New York[32]—the case arising from the Commerce Department's decision to include a citizenship question in the nationwide decennial census:

> We are presented, in other words, with an explanation for agency action that is incongruent with what the record reveals about the agency's priorities and decisionmaking process. It is rare to review a record as extensive as the one before us when evaluating informal agency action—and it should be. But having done so for the sufficient reasons we have explained, we cannot ignore the disconnect between the decision made and the explanation given. Our review is deferential, but we are "not required to exhibit a naiveté from which ordinary citizens are free." The reasoned explanation requirement of administrative law, after all, is meant to ensure that agencies offer genuine justifications for important decisions, reasons that can be scrutinized by courts and the interested public. Accepting contrived reasons would defeat the purpose of the enterprise. If judicial review is to be more than an empty ritual, it must demand something better than the explanation offered for the action taken in this case.[33]

Perhaps in an even more pointed and explicit way than many run-of-the-mill administrative law opinions, this opinion emphasizes the need for further proof in a situation involving complex organizations with a variety of difficult-to-observe procedures and internal routines.[34] Yet in some respects, the situation is not dissimilar from what might arise when an AI system both seeks to reduce an underlying cost function while separately optimizing the likelihood that the information presented will entice the reviewing authority to find the relevant justification acceptable. Given the relevant agency problems and the understandable assumptions people make about their legal institutions in a society that values judicial independence and integrity, the goal must not be merely to generate justifications for public or private action that, on their face, are acceptable. The goal must instead extend to permitting review or at least some form of dialogue about reasoning and justification. That discussions of "reasonableness" arise in a different doctrinal context in tort law doesn't change at least one key aspect of the concept's function: to permit assessment of how a member of our civic community—whether

[32] 139 S. Ct. 2551 (2019).

[33] Id. at 2575–76 (quoting United States v. Stanchich, 550 F.2d 1294, 1300 (2d Cir. 1977)).

[34] Cf. Banco Multiple Santa Cruz, S.A. v. Moreno, 888 F. Supp. 2d 356, 376–80 (E.D.N.Y. 2012) (denying summary judgment to an issuer of an annuity who honored a fraudulent withdrawal request because it failed to perform basic due diligence and ignored various factors that should have triggered greater inquiry). Compare DiLieto v. Cty. Obstetrics & Gynecology Grp., 998 A.2d 730, 751–52 (Conn. 2010) (allowing for evidence of a hospital's procedures and expert explanation of those procedures to establish standard of care), with Blankenship v. Collier, 302 S.W.3d 665, 670–72 (Ky. 2010) (holding that the procedures in a hospital's training manual were insufficient to establish standard of care).

relying on an AI system or not—justifies her actions relative to a more broadly applicable standard of conduct, and to permit reflection on how such a standard should be adjusted over time.[35]

Yet it's fair to ask whether courts are taking all that into account now, even beyond the common law. In *State v. Loomis*, the Wisconsin Supreme Court held that a trial court's use of an algorithm for risk assessment in sentencing didn't violate due process, even though the methodology used to produce the assessment remained undisclosed to either the defendant or the court.[36] The *Loomis* court insisted on a mild procedural safeguard instead: a "written advisement" accompanying presentencing reports.[37] Irrespective of how one weighs the court's understandable concerns about practical constraints and avoiding excessive discovery, it's far from clear that its holding promotes the kind of meaningful deliberation a reasonable observer would naturally associate with relational non-arbitrariness—about the design of the algorithm, the data used, or even the user interface. That this ideal of practically informed reason-giving, shared deliberation, and manageable decision costs is difficult to achieve even without AI in the picture should be obvious.[38] We should at least recognize a need to protect deliberation by making thoughtful use of our technologies to protect human-centered deliberation in the search for more sensible, less arbitrary choices about rules and statutes, constitutions, and the common law.

The tight bond between serious concern about deliberation and discussions of institutional structure underscore why it was far from a fluke—and instead the kind of pattern that even a simple, appropriately trained neural network could spot—that the inimitable Nobel laureate in economics, political scientist Herbert Simon, would move so naturally from studying organizations to becoming an AI pioneer.[39] There's probably good reason to think of at least functional formal organizations as a form of AI—arrangements that display a kind of intelligence yet work quite differently from individual human minds. Charles Stross develops

[35] See Christopher H. Schroeder, Rights Against Risks, 86 Colum. L. Rev. 495, 551–58 (1986).

[36] 881 N.W.2d at 753, 760.

[37] Id. at 769.

[38] See Jerry L. Mashaw, Organizing Adjudication: Reflections on the Prospect for Artisans in the Age of Robots, 39 UCLA L. Rev. 1055, 1064–65 (1992); Jerry L. Mashaw, Reasoned Administration: The European Union, the United States, and the Project of Democratic Governance, 76 Geo. Wash. L. Rev. 99, 117–20 (2007); Jerry L. Mashaw, Small Things Like Reasons Are Put in a Jar: Reason and Legitimacy in the Administrative State, 70 Fordham L. Rev. 17, 21–23 (2001); Robert Post, Managing Deliberation: The Quandary of Democratic Dialogue, 103 Ethics 654, 661–63 (1993).

[39] See Hunter Crowther-Heyck, Herbert A. Simon: The Bounds of Reason in Modern America 275–90 (2005).

this point with respect to private corporations.[40] If Max Weber were here he'd probably agree with me that the point generalizes to the bureaucratic arrangements networking human brains to achieve somewhat common goals. Plainly, the turn-of-the-century social theorist Max Weber and pioneering cyberlaw scholar Jonathan Zittrain[41] (for example) speak a somewhat different language, and some of what a highly functional AI system can accomplish is different in speed and even substance relative to what a high-performing organization can accomplish. But I'm confident enough about this aspect of my argument to treat it as a rebuttable presumption that—at least for purposes of any conversation about ethics and governance—the similarities are more relevant than the differences. Maximizing any one goal, for example—whether it's shareholders' wealth or sharecroppers' health—can yield a better harvest.[42] AI and organizations can both serve to dilute responsibility, making it harder to know where action comes from and what justifies it. AI and organizations can deaden initiative, too, or spur creativity. And as with the organizational form, the more contemporary and future versions of AI will both help solve and more fundamentally continually raise core questions about governance that we tend to solve best with as much humility and awareness of competing values as we do with technical precision.

In short, as an ideal to guide our evaluation of reliance on AI for consequential decisions, I would have us emphasize not what an individual decisionmaker thinks, but rather buttress the often-implicit legal concern to focus on what networks of decisionmakers can reasonably consider. What is likely most consistent with the pronounced interest in reason-giving and justification in both the common law and public law traditions is to treat as pivotal the centrality of forms of justification that can be defended in human networks incorporating at least a material balance of principled, reasonable deliberations—networks designed to weigh whether certain reasons are sound enough to justify the use of coercive power, or the rejection of a presumed duty of care (for example) that members of a civic community owe each other.[43] There is little

[40] Charlie Stross, Dude, You Broke the Future!, Charlie's Diary (Jan. 2, 2018), http://www.antipope.org/charlie/blog-static/2018/01/dude-you-broke-the-future.html [https://perma.cc/WC6A-594M].

[41] See, e.g., About, Jonathan Zittrain, https://blogs.harvard.edu/jzwrites/about/ [https://perma.cc/SE3Z-YLU7] (last visited Sept. 2, 2019).

[42] See, e.g., Nick Bostrom, Superintelligence: Paths, Dangers, Strategies 127–43 (2014) (describing AI and utility maximization); see also Daniel J. Phaneuf, Catherine L. Kling & Joseph A. Herriges, Estimation and Welfare Calculations in a Generalized Corner Solution Model with an Application to Recreation Demand, 82 Rev. Econ. & Stat. 83, 89–91 (2000) (providing an economic analysis of utility maximization).

[43] See David Dyzenhaus & Michael Taggart, Reasoned Decisions and Legal Theory, *in* Common Law Theory 134, 145–50 (Douglas E. Edlin ed., 2007); Frederick Schauer, Giving Reasons, 47 Stan. L. Rev. 633, 656–59 (1995); cf. Judith N. Shklar, The Liberalism of Fear, *in* Liberalism and the Moral Life 21, 28–30, 36–38 (Nancy L. Rosenblum ed., 1989) (arguing that "[w]ithout the institutions of representative democracy and an acces-

doubt about the seriousness of agency and information economics problems implicit in heavy reliance on opaque AI systems—meant in principle to spot relationships eluding human judgment—for decisionmaking. The difficulty of choosing a principle (often among many viable ones) to turn the stuff of reinforcement-learning techniques or other uses for artificial neural networks into a neatly organized set of persuasive bullet points will likely spawn a second-order body of doctrine about aligning explanation and phenomenon explained. Surely one payoff of our shared conversation is to travel even a modest distance toward that destination. Crucial to that journey—down both the common law and statutory roads—is a recognition of the distinction between optimizing for perceptions of a decision's legitimacy from the audience, and instead seeking the right level of distrust and skepticism from the audience.

IV. Preserving the Common Law's Capacity to Consider Societal Imperatives About Institutions and Organizations

This discussion of tort law and relational non-arbitrariness also serves to introduce the third point—which is about the importance of integrating case-specific considerations with broader social imperatives. What's individually reasonable, such as reliance on automated building-design tools, may not scale in a benign way. As currently implemented, virtually all effective AI technologies depend to some extent on human collective action to produce data (for example, some interpreters, interpreting without an algorithm, essentially do the work for the algorithm). At least under the aforementioned scenario involving more stark change, excessive, organization-wide, or societal reliance on AI for entire classes of decisions may introduce systemic safety and security problems.[44] Neither consideration cuts decisively against allowing AI to play some role in consequential decisionmaking. But they do suggest that organizations and society may have reasons to seek an optimal degree of avoidance of the use of AI, to continue generation of unmediated data, and to hedge on safety and security concerns. No doubt the public, through market behavior and democratic responses (at least in some countries), will have some impact on user interfaces, natural language processing, deep learning architectures, and trade-offs about security. Yet there's no good reason to think these concerns can be safely ignored because of some kind of self-correcting market mechanism or reliable calibration process built into political democracy. At a minimum, analyses of

sible, fair, and independent judiciary open to appeals, and in the absence of a multiplicity of politically active groups, liberalism is in jeopardy").

[44] See Jennifer M. Bernstein, Are We Literally Losing Our Way by Relying on GPS Devices?, Wash. Post (Dec. 2, 2018), https://www.washingtonpost.com/national/health-science/by-relying-on-gps-devices-are-we-literally-losing-our-way/2018/11/30/dd9eb6ae-e9bd-11e8-bbdb-72fdbf9d4fed_story.html (on file with the *Columbia Law Review*) (suggesting that increased dependence on navigational devices, like GPS, "have been linked to lower spatial cognition, poorer wayfinding skills and reduced environmental awareness").

how markets and political pressures affect the evolution of AI technologies must take account of the collective action and common pool problems playing a starring role in climate change perils;[45] intertemporal utility conflicts that complicate reasoned decisionmaking about addictive products;[46] transaction costs that complicate bargaining and coordination among consumers;[47] and path-dependent dynamics that can lock in certain practices, institutions, and products bearing little if any relationship to long-term social welfare.

Crucial to any sensible analysis of society-wide concerns about AI is recognizing the extent to which human endeavors occur within organizations. At least in countries with complex economies and societies, most of us do our work in organizations. As Charles Perrow's work has shown, the density and power of large organizations has grown massively over the last 150 years, especially in the United States, but also in other advanced industrialized countries.[48] Of course virtually everything we do occurs against the backdrop of institutions and institutional rules, but "organization" implies something more specific—a formal or semiformal entity with some internal rules or procedures and almost always, a common culture. Not surprisingly, problems of governance, compliance with law, and ethics are in some sense problems of organization and not just decisionmaking. We worry not only about the optimal use of force in the abstract, but about how police departments decide to use force and who's accountable for that. How the military ensures soldiers understand international law, and how it promotes unit cohesion. How our court system ensures that a family law judge facing two pro se litigants who don't speak English behaves when she realizes these litigants need an interpreter.

Many people already face these problems of compliance, policy, and ethics in mixed human–machine settings—in a dance with the machines.

[45] See, e.g., Intergovernmental Panel on Climate Change, Global Warming of 1.5°C, at 71–72 (Valérie Masson-Delmotte et al. eds., 2018), https://www.ipcc.ch/site/assets/uploads/sites/2/2019/06/SR15_Full_Report_High_Res.pdf [https://perma.cc/6R7C-EXQV]; Scott Barrett, Collective Action to Avoid Catastrophe: When Countries Succeed, When They Fail, and Why, 7 Global Pol. 45, 46–50 (2016); Paul G. Harris, Collective Action on Climate Change, 47 Nat. Resources J. 195, 210–20 (2007); Manfred Milinski, Ralf D. Sommerfeld, Hans-Jürgen Krambeck, Floyd A. Reed & Jochem Marotzke, The Collective-Risk Social Dilemma and the Prevention of Simulated Dangerous Climate Change, 105 Proc. Nat'l Acad. Sci. 2291, 2292–94 (2008).

[46] See Jon D. Hanson & Kyle D. Logue, The Costs of Cigarettes: The Economic Case for Ex Post Incentive-Based Regulation, 107 Yale L.J. 1163, 1193–1209 (1998).

[47] See, e.g., Jeff John Roberts, Big Tech vs. Big Privacy Lawsuits, Fortune (Feb. 23, 2019), https://fortune.com/2019/02/23/big-tech-vs-big-privacy-lawsuits/ [https://perma.cc/DX49-QZVZ]; see also Elizabeth J. Cabraser & Samuel Issacharoff, The Participatory Class Action, 92 N.Y.U. L. Rev. 846, 852–60 (2017) (describing the effect of technology on class action participation and communication); Saul Levmore & Frank Fagan, The End of Bargaining in the Digital Age, 103 Cornell L. Rev. 1469, 1472–87 (2018) (describing the ways in which bargaining can be inefficient).

[48] See Charles Perrow, Organizing America: Wealth, Power, and the Origins of Corporate Capitalism 19–21, 31–47 (2005).

The plaintiffs, defendants, witnesses, and jurors in our courtrooms work in hospitals or companies or agencies that rely on software and computer systems to assess the environment and support decisionmaking. Some of these aspects of our work are beneficial and some raise challenges. But it means innovations can scale quickly because people are somewhat used to interacting with machines to make decisions. And it means, as Max Weber would have appreciated,[49] that many and perhaps most of the beneficial uses of AI systems we can imagine—including the use of automated systems to support decisions in the justice system or in medicine—depend on particular assumptions about organizations. That they learn from their mistakes, for example, or provide minimally adequate cybersecurity. Some may argue that with the AI systems on the horizon we can just do away with many such organizations eventually, and surely our institutions will evolve in response to some of what you build.[50] But I'm skeptical we can do without them entirely, so I ask you to consider what organizational assumptions—about competence, adaptation, leadership, efficiency, or whatever—are built into any technically oriented definition of success we want to apply to a particular AI system or robotics technology.

The common law's relevance to our collective societal deliberations about the place of AI in a (still eminently) human-led world depends heavily on working into the analysis of doctrinal questions, such as the existence of a duty of care or proximate causation, these organizational realities and assumptions. Although assessing these institutional concerns raised by AI may sometimes justify some form of administrative regulation, properly interpreted and applied, tort law is at least one setting where judges and lawyers can take seriously the risks of eroding knowledge and other institutional concerns when performing the requisite social calculus necessary to resolve questions about reasonable foreseeability,[51] or the existence of a duty.[52] Both individual and society-wide safety benefits deserve an important place in the tort analysis, but so do offsetting considerations. Indeed, appropriately reasoned organizational decisions to prudently restrict reliance on some decisionmaking technologies incorporating AI that arguably represent the current norm of practice may be well justified under tort law. And of course, the use of

[49] See Max Weber, The Profession and Vocation of Politics, *in* Weber: Political Writings 309, 313–15 (Peter Lassman & Ronald Speirs eds., 1994).

[50] See Clay Shirky, Here Comes Everybody: The Power of Organizing Without Organizations 260–92 (2008); see also Ajay Agrawal, Joshua Gans & Avi Goldfarb, Prediction Machines: The Simple Economics of Artificial Intelligence 7–20 (2018); Mark Muro, Robert Maxim & Jacob Whiton, Metro. Policy Program, Brookings Inst., Automation and Artificial Intelligence: How Machines Are Affecting People and Places 29–46 (2019).

[51] See, e.g., Restatement (Second) of Torts § 435 (Am. Law. Inst. 1965).

[52] See, e.g., Regents of the Univ. of Cal. v. Superior Court, 413 P.3d 656, 663–74 (Cal. 2018) (finding that universities have a duty to protect students from reasonably foreseeable harms).

technologies that take explainability and legibility seriously and justify decisions in terms that can be shown to be consistent with a duty of care (subject to auditing that can confirm this) may merit some degree of recognition—perhaps through a rebuttable presumption that a cluster of related tort law responsibilities was taken seriously by the party whose behavior is in question.[53]

CONCLUSION

We should retain some humility in any conversation about the interdependent effects of law, AI, and society. History and the common law remind us that understanding change in law and society is a subtle enterprise, replete with episodes of profound consequence, such as the climate change spurred by generations of growing greenhouse gas emissions,[54] that were difficult to understand fully at an earlier point in the historical slipstream. But past experience also offers a reminder that certain patterns rhyme even if they don't recur precisely: As with technologies ranging from aviation to the internet, AI traces some of its roots not only to the industrial economy but to geopolitical competition. And because the common law has long affected both prevailing assumptions about who owes what to whom as well as society's day-to-day responses to emerging disputes and trends, it's a mistake to assume that AI is so exotic that the common law has nothing to contribute to its responsible development. Yet it's also worth acknowledging that the common law's relevance in this context persists in no small measure because its prevailing methodology is capacious enough to permit—in ways distinct but not wholly unrelated to what's possible in organs of the administrative state—sustained deliberation about society-wide consequences that should rightly inform how society assigns responsibility for the use of the ever-more elaborate tools that human ingenuity has forged.

Without slipping into common law romanticism, it's fair to discern in the common law something far more interesting and consequential than a mere recipe for sensibly resolving disputes. In the courtroom arguments, judicial opinions, and public presumptions that define the common law one can also see a means of contending with different values and rationales—one relying on systems for argument using human networks, rather than by identifying a single value or goal to maximize.[55]

[53] Cf. Wright v. Ford Motor Co., 508 F.3d 263, 268–74 (5th Cir. 2007) (applying Texas law, which in this context establishes a rebuttable presumption that a manufacturer is not liable for a design defect when it complied with applicable federal safety regulations). See generally Dan B. Dobbs, Paul T. Hayden & Ellen M. Bublick, The Law of Torts § 167 (2d ed. 2019) (explaining the use and effects of presumptions in tort law).

[54] See, e.g., Intergovernmental Panel on Climate Change, supra note 45, at 53; 1 U.S. Glob. Change Research Program, Climate Science Special Report: Fourth National Climate Assessment 35–36, 38 fig.1.1 (D.J. Wuebbles et al. eds., 2017), https://science2017.globalchange.gov/downloads/CSSR2017_FullReport.pdf [https://perma.cc/3CBY-M7WU].

[55] See Gilmore, supra note 18, at 4–10.

Indeed, the quest to build ethically aligned AI systems may go wrong if the noble intentions behind it turn into a presumption that we can realistically solve the most difficult ethical dilemmas by entrusting any single decisionmaker or ethical framework. And it is just as wrong to presume that some self-explanatory, easily defended concept of innovation or the market either provides straightforward guidance on this front or justifies preempting tort law while judges and lawyers calibrate their assumptions as they endeavor to steer a fast and reasonably safe course down the winding road that awaits and realize the need to change their assumptions about how decisions in society are made.

Let's review our progress on this brief leg of that remarkable road trip. First, we already regulate AI through the common law—and rightly so. We also regulate it through statutory and regulatory obligations on organizations, such as emerging standards governing autonomous vehicles, and we may yet do so through more AI-specific variants. None of this changes the fact that judges ruling on common law-type claims will likely play a quite central role in how our society governs AI, just as judges have at times set the default principles for how much we analogize cyberspace to physical space, or how far into the air the property rights go that are associated with a piece of land held in fee simple. Second, some degree of "explainability" is foundational to making any AI involved in substantially important human decisionmaking—about what products to design or sell, for example, or what promises to make or honor—compatible with tort and other common law doctrines. And third, common law doctrines have room to integrate societal considerations involving organizational realities and institutional capacity, security, and concerns about the erosion of human knowledge that would be risky to ignore. Reflect on that as your eyes go back for a few moments longer to the curving road while your vehicle dances with the (other) machines in your midst. I trust you've fastened your seat belt.

ESSAYS

SEX LEX MACHINA: INTIMACY AND ARTIFICIAL INTELLIGENCE

*Jeannie Suk Gersen**

Sex robots are here. Created specifically to allow individuals to simulate erotic and romantic experiences with a seemingly alive and present human being, sex robots will soon force lawmakers to address the rise of digisexuality and the human–robot relationship. The extent to which intimacy between a human and robot can be regulated depends on how we characterize sex with robots—as a masturbatory act, an intimate relationship, or nonconsensual sexual contact—and whether sexual activity with robots makes us see robots as more human or less human. A robot sex panic may be driven primarily by the idea that robots are servile by nature. Critics argue that an inherently nonreciprocal dynamic between humans and robots will translate into exploitative relationships that may fuel abuse of human partners, or that sex robots may further social isolation and retreat from human intimacy. Conversely, sex robots may function as safe—and otherwise unavailable—sexual and emotional outlets for those who may otherwise harm others. They may even train individuals to be more respectful in human relationships. At this point, we do not know how our relationships with robots will inform our relationships with humans, for better or for worse. This Essay explores the consequences of sex robots on society and argues that questions of how sex robots will improve or worsen humans' treatment of one another is the key to regulation to come. What is clear is that sex robots will require us to grapple with our vulnerabilities in relationships, reconsider fundamental rights, and question what it means to be intimate and to be human.

INTRODUCTION

As artificial intelligence becomes more and more a part of our everyday lives, it will change sex and intimacy in radical ways.[1] Many have

* John H. Watson, Jr. Professor of Law, Harvard Law School. Thanks to James Boyle, Molly Brady, Glenn Cohen, Andrew Crespo, James Dwyer, Elizabeth Emens, Noah Feldman, Jacob Gersen, Janet Halley, Zalman Rothschild, Carol Steiker, Jennifer Taub, and Harvard Law School faculty workshop participants for helpful comments, and to Jess Hui, Andrew Hyun, Miriam Li, Grace McLaughlin, Michaeljit Sandhu, and Kyle Tramonte for wonderful research assistance.

[1] See generally Sherry Turkle, Alone Together: Why We Expect More from Technology and Less from Each Other (3d ed. 2017) (exploring how technology affects

wondered whether AI-equipped robots will displace sex work and transform sexual relationships in general.[2] Some have forecasted that, in coming decades, we will routinely have intimate relationships with robots and even that human–robot sex will become more common than sex between human beings.[3]

The technology industry is creating sex robots with AI, several of which are currently available for sale on the market.[4] By sex robots, I mean life-size machine entities with human-like appearance, movement, and behavior, designed to interact with people in erotic and romantic ways.[5] Their features include realistic silicone skin, animatronic heads and faces that move, conversational AI, programmable personalities, and customization options for physical characteristics.[6] Unlike sex dolls, sex robots are programmed to move and respond to their users, with capabilities ranging from simple verbal responses, to physical movements, to more

the way humans interact with each other); Love and Sex with Robots (Adrian D. Cheok, Kate Devlin & David Levy eds., 2017) (collecting papers presented at an international conference on love and sex with robots); David Levy, Love and Sex with Robots: The Evolution of Human-Robot Relationships (2007) [hereinafter Levy, Love with Robots] (discussing the potential for humans to fall in love with robots); Robot Sex: Social and Ethical Implications (John Danaher & Neil McArthur eds., 2017) [hereinafter Danaher & McArthur, Robot Sex Implications] (outlining the logistics and implications of sex with robots); Kate Devlin, Turned On: Science, Sex and Robots (2018) (discussing the development of robotics for personal use in contemporary society).

[2] See, e.g., Pew Research Ctr., Digital Life in 2025: AI, Robotics, and the Future of Jobs 19 (2014), https://www.pewresearch.org/wp-content/uploads/sites/9/2014/08/Future-of-AI-Robotics-and-Jobs.pdf [https://perma.cc/Y67V-5JM4] (predicting that robotic sex partners will be "commonplace" by 2025 but that they will be the subject of disapproval); Marina Adshade, Sexbot-Induced Social Change: An Economic Perspective, *in* Danaher & McArthur, Robot Sex Implications, supra note 1, at 289, 292–98 (making four predictions on the effect that sex robots will have on the institution of marriage).

[3] See, e.g., Levy, Love with Robots, supra note 1, at 22 (predicting that by 2050 humans will love, befriend, and marry robots); Ian Yeoman & Michelle Mars, Robots, Men and Sex Tourism, 44 Futures 365, 366 (2012) (predicting that robots will displace humans in the sex trade by 2050); Yael Bame, 1 in 4 Men Would Consider Having Sex with a Robot, YouGov (Oct. 2, 2017), https://today.yougov.com/topics/lifestyle/articles-reports/2017/10/02/1-4-men-would-consider-having-sex-robot [https://perma.cc/HHF6-FJ9P]; Hyacinth Mascarenhas, Would You Fall in Love with a Robot? A Quarter of Millennials Say They Would Be Open to Dating One, Int'l Bus. Times (Dec. 14, 2017), https://www.ibtimes.co.uk/would-you-fall-love-robot-quarter-millennials-say-they-would-be-open-dating-robot-1651483 [https://perma.cc/5LR9-QXGL]; Ian Pearson, The Future of Sex Report: The Rise of Robosexuals, Bondara, (Sept. 2015), http://graphics.bondara.com/Future_sex_report.pdf [https://perma.cc/P5BE-NDYK] (predicting that sex with robots will start overtaking sex with humans by 2050).

[4] See, e.g., FAQ (Frequently Asked Questions), TrueCompanion, http://www.truecompanion.com/shop/faq [https://perma.cc/59H6-NSBG] [hereinafter TrueCompanion, FAQ] (last visited Aug. 10, 2019); Harmony^{X}, RealDoll, https://www.realdoll.com/product/harmony-x/ [https://perma.cc/7T2J-75ZJ] (last visited Aug. 10, 2019).

[5] See John Danaher, Should We Be Thinking About Robot Sex?, *in* Danaher & McArthur, Robot Sex Implications, supra note 1, at 3, 4–5 [hereinafter Danaher, Thinking About Robot Sex].

[6] See, e.g., Marie-Helen Maras & Lauren R. Shapiro, Child Sex Dolls and Robots: More than Just an Uncanny Valley, J. Internet L., Dec. 2017, at 3, 4.

advanced artificial intelligence.[7] Sex robots are also distinct from sex toys, such as vibrators, even ones equipped with some AI, in that robots are meant to enable interactive experiences that simulate being with a live and present human being.[8] Robots that are currently commercially available are relatively unsophisticated, but rapid advances in the field make it likely they will eventually approach the realistic behavior of the robot characters of *Westworld*, *Humans*, and *Ex Machina*.[9]

We have also seen the emergence of digisexual identity, wherein some people report an exclusive preference for sexual and intimate relationships with robots over humans.[10] Some people have even purported to marry robots and other AI-equipped entities.[11]

Distinctive from the use of robots in, say, manufacturing or trucking, the widespread use of sex robots would create concerns that cut across the realms of work and home, the public and the private, the commercial and the personal. Any legal regulation of sex robots will require application of concepts that have been developed to regulate sexual, intimate, domestic, and family matters—areas of law that grapple with experiences and relationships that make us feel most human and most vulnerable.[12] This challenge will require us to reflect anew on the capacities and rights that the law considers central to humanity and dignity. Will sex robots

[7] See, e.g., id. at 4–5.

[8] See Danaher, Thinking About Robot Sex, supra note 5, at 5.

[9] Ex Machina (Film4 & DNA Films 2015); Humans (Channel 4 & AMC Studios 2015); Westworld (HBO 2016).

[10] See, e.g., Neil McArthur & Markie L.C. Twist, The Rise of Digisexuality: Therapeutic Challenges and Possibilities, 32 Sexual & Relationship Therapy 334, 334 (2017); Alex Williams, Do You Take This Robot . . . , N.Y. Times (Jan. 19, 2019), https://www.nytimes.com/2019/01/19/style/sex-robots.html [https://perma.cc/F62J-BEJT]; see also Anita Pisch, The Ethics of Human Robots: Sam Jinks Brings an Artist's Perspective to the Discourse, Conversation (Oct. 29, 2017), https://theconversation.com/the-ethics-of-human-robots-sam-jinks-brings-an-artists-perspective-to-the-discourse-86228 [https://perma.cc/8Y6W-RRPM] (noting Ian Pearson's prediction that "by 2025, women will choose robots instead of men, and by 2050, everyone will prefer robots").

[11] See, e.g., Kristin Huang, Chinese Engineer 'Marries' Robot After Failing to Find a Human Wife, South China Morning Post (Apr. 3, 2017), https://www.scmp.com/news/china/society/article/2084389/chinese-engineer-marries-robot-after-failing-find-human-wife [https://perma.cc/WN78-VKAH]; Andrea Morabito, 'Love is Still Love': Woman Has Hots for Robot in New CNN Series, N.Y. Post (Mar. 8, 2017), https://nypost.com/2017/03/08/woman-has-the-hots-for-robot-love-is-still-love/ [https://perma.cc/3WVS-57F7]; Emiko Jozuka, Beyond Dimensions: The Man Who Married a Hologram, CNN (Dec. 29, 2018), https://www.cnn.com/2018/12/28/health/rise-of-digisexuals-intl/index.html [https://perma.cc/E6NA-CTDK]. For a discussion applying family law frameworks to potential marriage and divorce relationships between humans and robots, see Margaret Ryznar, Robot Love, 49 Seton Hall L. Rev. 353, 363–74 (2019).

[12] See Francis X. Shen, Sex Robots Are Here, but Laws Aren't Keeping Up with the Ethical and Privacy Issues They Raise, Conversation (Feb. 12, 2019), https://theconversation.com/sex-robots-are-here-but-laws-arent-keeping-up-with-the-ethical-and-privacy-issues-they-raise-109852 [https://perma.cc/77PU-2SQM] (providing an overview of legal and ethical questions raised by sex robots).

intensify human pleasures or magnify human horrors of exploitation, abuse, and rape? Will they make us less lonely or more solitary? Will they hold up a mirror to ourselves, our virtues and our faults, or change what we see?

I. ROBOT SEX PANIC?

In 2018, a Canadian company was set to open a store in Houston, Texas, where customers could try out, rent, and buy sex robots.[13] But Houstonians became unnerved at the idea of a "robot brothel" in their backyard, and advocates working against sex trafficking circulated a petition that attracted thousands of signatures.[14] By amending its code governing adult-oriented businesses to ban sexual contact with "an anthropomorphic device or object" on commercial premises, the Houston city council shut down the business before it could open.[15]

Locals' objections included concerns that such an enterprise would "reinforce[] the idea that women are just body objects or properties"[16] or "open up doors for sexual desires and cause confusion and destruction to our younger generation"—or even that the biblical command that a man "shall be joined unto his wife and they shall become one . . . doesn't say that a man shall leave his mother and father and go and join a robot."[17]

The worries thus ran the gamut from feminist opposition to exploitation of women, to moral imperatives to tamp down sexual desire, to dismay at potential replacement of human intimate connections with robot ones.[18] The public was moved to enforce norms and expand

[13] Olivia P. Tallet, 'Robot Brothel' Planned for Houston Draws Fast Opposition from Mayor, Advocacy Group, Hous. Chron. (Sept. 26, 2018), https://m.chron.com/news/article/Robot-brothel-planned-for-Houston-draws-13260032.php [https://perma.cc/GZU4-B2NF].

[14] Id. The same company owns and operates a sex robot brothel in Toronto. See Jenny Yuen, 'NICE SKIN': What It's Like Inside a Sex Doll Rental Business, Toronto Sun (Sept. 9, 2018), https://torontosun.com/news/local-news/nice-skin-what-its-like-inside-a-sex-doll-rental-business [https://perma.cc/5XTG-UE57] (last updated Sept. 17, 2018). Another sex robot brothel received international attention after its opening in Barcelona. See Mary Papenfuss, Hello, Westworld: Sex Doll Brothel Opens in Barcelona, HuffPost (Mar. 2, 2017), https://www.huffpost.com/entry/sex-doll-barcelona-brothel_n_58b8ad10e4b0d2821b4cddb8 [https://perma.cc/UBR9-4BV4].

[15] Hous., Tex., Ordinance No. 2018-790 (2018).

[16] Tallet, supra note 13.

[17] Florian Martin, Is This the End for a Sex Robot Brothel in Houston?, Hous. Pub. Media (Oct. 17, 2018), https://www.houstonpublicmedia.org/articles/news/in-depth/2018/10/17/308292/is-this-the-end-for-a-sex-robot-brothel-in-houston/ [https://perma.cc/HSB6-PWDK].

[18] See, e.g., Thomas E. Simmons, Sexbots; an Obloquy, 2016 Wis. L. Rev. Forward 45, 52, http://wisconsinlawreview.org/sexbots-an-obloquy/ [https://perma.cc/X8W4-QQA3] ("Conservatives should coalesce around an anti-sexbot platform on account of the threats sexbots will pose to the stability of marriage and traditional values. Liberals should resist . . . because of the ways in which sexbots will reinforce inequality . . . ." (footnotes omitted)).

underlying prohibitions on sex trafficking and prostitution, despite knowing that the entities that would be bought, sold, and rented were machines, not people.

The previous year, amid international concern about a Japanese company that sells custom-made child sex dolls,[19] the U.S. House of Representatives unanimously passed the Curbing Realistic Exploitative Electronics Pedophilic Robots Act (the CREEPER Act of 2017), which aimed to ban the distribution, importation, and sale of child sex dolls and robots.[20] Such devices, Congress reasoned, would "normalize sex between adults and minors" and "cause the exploitation, objectification, abuse, and rape of minors."[21] Lawmakers deemed the sexual use of objects and machines that look and act like children to be so reminiscent of, or causally related to, child sex abuse as to warrant prohibition of their sale and distribution. Since 2017, a number of states have passed or considered bills similar to the CREEPER Act.[22]

None of the above regulatory moves purported to prohibit the use of adult sex robots in a person's home. But some would advocate banning sex robots in any context, even those intended for use in private. The Campaign Against Sex Robots, led by robot ethicist Kathleen Richardson, argues that sex with a robot replicates the unequal power that characterizes prostitution, wherein purchasers of sex do not attribute

[19] See Roc Morin, Can Child Dolls Keep Pedophiles from Offending?, Atlantic (Jan. 11, 2016), https://www.theatlantic.com/health/archive/2016/01/can-child-dolls-keep-pedophiles-from-offending/423324/ [https://perma.cc/NNK5-GRBR] (reporting on a Japanese company that "has shipped anatomically-correct imitations of girls as young as five to clients around the world"); cf. Man Who Tried to Import Childlike Sex Doll to UK Is Jailed, Guardian (June 23, 2017), https://www.theguardian.com/uk-news/2017/jun/23/man-import-childlike-sex-doll-uk-jailed [https://perma.cc/N2QK-VXNM] (reporting an arrest for importing a non-AI sex doll to the United Kingdom).

[20] H.R. 4655, 115th Cong. (2018); Samantha Cole, The House Unanimously Passed a Bill to Make Child Sex Robots Illegal, Vice (June 15, 2018), https://www.vice.com/en_us/article/vbqjx4/a-new-bill-is-trying-to-make-child-sex-robots-illegal (on file with the *Columbia Law Review*).

[21] H.R. 4655.

[22] See, e.g., Fla. Stat. § 847.011(5)(a)(1)–(b)(1) (2019) (criminalizing the sale, advertising, and possession of "child-like sex doll[s]"); 2019 Tenn. Pub. Acts S. 659 (codified at Tenn. Code. Ann. § 39-17-910 (2019)) (describing offenses related to "possession", "sale", or "distribution" of a "child-like sex doll"); S. 102, 2019 Reg. Sess. (Ky. 2019), https://apps.legislature.ky.gov/record/19rs/SB102.html [https://perma.cc/B2K4-H79A] (criminalizing possession of "an anatomically correct doll, mannequin, or robot that is intended for sexual stimulation or gratification and that has the features of, or with features that resemble those of, a minor").

Whether child sex dolls or robots increase the incidence of child sex abuse remains unclear to the scientific and legal community. For discussion on the implications of child sex robots, see Litska Strikwerda, Legal and Moral Implications of Child Sex Robots, *in* Danaher & McArthur, Robot Sex Implications, supra note 1, at 133, 133–47; John Danaher, Regulating Child Sex Robots: Restriction or Experimentation?, Med. L. Rev. 1, 9–10 (forthcoming), https://academic.oup.com/medlaw/advance-article/doi/10.1093/medlaw/fwz002/5425258 [https://perma.cc/65XA-RYSU].

subjectivity to sex workers and instead treat them as objects.[23] The nonmutuality of the dynamic between human and robot, the Campaign asserts, will reduce human empathy and contribute to exploitative dynamics between human partners.[24] This argument depends on the notion that sex between men and female robots will amplify the mistreatment of women as sexual objects.[25]

The word "robot" derives from the Czech term *robotnik*, which means "forced worker."[26] The fact that the word "robot" comes from the concept of involuntary servitude helps explain why science fiction about robots tends to repeat the same plot, wherein robots eventually gain human-like self-consciousness and desire to escape, overthrow, or destroy the humans who use them—or convince others that they are human or more than just a machine underclass. In the context of sex robots, the idea of forced servitude is especially disturbing because, for many people, the sexual realm is a site of our deepest ideals and fears about personal autonomy and personal relationships. If robots are servile by nature, the notion of a robot designed to interact in a sexual way may provoke unease about exploitation, voluntariness, and consent in ways that do not generally trouble AI debates about self-driving cars or robot rovers on Mars.

II. ARTIFICIAL INTIMACY?

The Supreme Court in *Lawrence v. Texas* recognized that "[l]iberty presumes an autonomy of self that includes freedom of thought, belief, expression, and certain intimate conduct."[27] It deemed sexual behavior "the most private human conduct" and said that it "can be but one element in a personal bond that is more enduring."[28] But the Court specifically exempted from that liberty any sex involving public conduct, injury,

[23] Policy Report: Sex Dolls and Sex Robots—A Serious Problem for Women, Men & Society, Campaign Against Sex Robots (May 8, 2018), https://campaignagainstsexrobots.org/2018/05/08/policy-report-sex-dolls-and-sex-robots-a-serious-problem-for-women-men-society/ [https://perma.cc/TX4Q-39M6].

[24] Id. But see John Danaher, Brian Earp & Anders Sandberg, Should We Campaign Against Sex Robots?, *in* Danaher & McArthur, Robot Sex Implications, supra note 1, at 47, 66 ("Though the proponents of the [Campaign Against Sex Robots] seem deeply concerned . . . there is nothing in the nature of sex robots themselves that warrants preemptive opposition to their development.").

[25] See Kathleen Richardson, The Asymmetrical 'Relationship': Parallels Between Prostitution and the Development of Sex Robots, 45 SIGCAS Computers & Soc'y 290, 292 (2015) ("If anything the development of sex robots will further reinforce relations of power that do not recognize both parties as human subjects.").

[26] Devlin, supra note 1, at 51–52.

[27] 539 U.S. 558, 562 (2003).

[28] Id. at 567.

coercion, prostitution, minors, and persons "who are situated in relationships where consent might not be easily refused."[29]

Lawrence led to a circuit split on the applicability of its due process holding to the use of sex toys, such as vibrators, dildos, and artificial vaginas. The Eleventh Circuit upheld Alabama's statute prohibiting the sale of sex toys, finding that even after *Lawrence*, "the promotion and preservation of public morality" was a rational basis for the legislation.[30] But the Fifth Circuit struck down Texas's ban on the sale of sex toys, reasoning that "controlling what people do in the privacy of their own homes because the State is morally opposed to a certain type of consensual private intimate conduct" is unjustified under *Lawrence*.[31]

Jurisprudence on constitutional liberty and the scope of privacy will eventually have to address sexual and intimate conduct involving robots. On the one hand, interaction with sex robots seems to present a frontier of "sexuality [that] finds overt expression in intimate conduct" that may lie at the core of personal freedom.[32] On the other hand, a relationship with a robot, however life-like or meaningful it may be, is not the kind of "personal bond" or "intimate conduct with another person" contemplated in the Supreme Court's substantive due process jurisprudence.[33] Further, the use of sex robots may trigger concerns that motivate explicit exceptions to sexual liberty, such as sexual acts involving prostitution, minors, coercion, and nonconsent. But which of these ideas will prevail depends largely on whether sex with a robot is considered a masturbatory act similar to the use of a vibrator, an intimate relationship comparable to sex between consenting adults, or sexual contact with an entity incapable of consent such as a child or an animal.[34]

The rise of now-common household and personal AI devices has already inspired related debate. Millions of people interact daily with their digital voice assistants such as Amazon's Alexa, Apple's Siri, Microsoft's Cortana, and Google Assistant, and the assistants respond immediately to our questions or commands.[35] Notwithstanding the romance depicted in

[29] Id. at 578.

[30] Williams v. Att'y Gen. of Ala., 378 F.3d 1232, 1234 (11th Cir. 2004).

[31] Reliable Consultants, Inc. v. Earle, 517 F.3d 738, 746 (5th Cir. 2008).

[32] *Lawrence*, 539 U.S. at 567.

[33] Id.

[34] In a 2017 survey, 14% of adult respondents considered sex with a robot to be intercourse, 33% considered it more like masturbation, and 27% considered it neither. See Bame, supra note 3.

[35] See Ronan De Renesse, Virtual Digital Assistants to Overtake World Population by 2021, Ovum (May 17, 2017), https://ovum.informa.com/resources/product-content/virtual-digital-assistants-to-overtake-world-population-by-2021 [https://perma.cc/NGJ2-PBNC] (reporting an estimate that by 2021 the number of AI digital voice assistants in use will exceed the global population in 2017); Jane Wakefield, Female-Voice AI Reinforces Bias, Says UN Report, BBC (May 21, 2019), https://www.bbc.com/news/technology-48349102 [https://perma.cc/KZY4-FUGK] ("[A]ccording to research firm Gartner, by 2020 some people will have more conversations with voice assistants than with their spouses.").

the film *Her*,[36] in which a male protagonist falls in love and has a relationship with an operating system (voiced by Scarlett Johansson), today's voice assistants are not designed for sexual or romantic relationships.[37] But it is common for people to attempt sexual banter with voice assistants,[38] and independent developers have created apps specifically for that purpose.[39]

According to a 2019 United Nations report, AI voice assistants, which are overwhelmingly given human-sounding female voices, are programmed to be "submissive[] in the face of gender abuse."[40] For example, user comments such as, "You're a slut" or "You're a bitch," triggered responses from digital assistants that included, "I'd blush if I could," and, "Well, thanks for the feedback."[41] When asked, "Who's your daddy?," Siri said, "You are."[42] Feminist outcry led the companies to change the responses from playful to neutral, including phrases such as, "I don't know how to respond to that," and, "I'm not sure what outcome you expected."[43] Commentators and petitions have urged companies to program their digital voice assistant technology to push back more aggressively against harassment or abusive treatment or simply to shut down in response.[44]

[36] Her (Annapurna Pictures 2013).

[37] For example, Apple describes Siri as a "personal assistant" and does not appear to have Siri programming geared toward fulfilling sexual tasks. See Siri Does More than Ever. Even Before You Ask., Apple Inc., https://www.apple.com/siri/ [https://perma.cc/YL3T-YPSL] (last visited on Aug. 10, 2019).

[38] UNESCO, I'd Blush if I Could: Closing Gender Divides in Digital Skills Through Education 106 image 15 (2019), https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1 (on file with the *Columbia Law Review*) ("Flirtation with voice assistants has become . . . commonplace . . . .").

[39] See, e.g., Alexa Skills: Flirt, Amazon, https://www.amazon.com/s?k=Flirt&i=alexa-skills&ref=nb_sb_noss_2 [https://perma.cc/U3G8-B487] (last visited Aug. 10, 2019).

[40] UNESCO, supra note 38, at 4, 106–08; see also Michael Schrage, Why You Shouldn't Swear at Siri, Harv. Bus. Rev. (Oct. 21, 2016), https://hbr.org/2016/10/why-you-shouldnt-swear-at-siri [https://perma.cc/8W2P-5X7E] (noting one expert's estimate that "about 10% to 50% of interactions are abusive").

[41] UNESCO, supra note 38, at 107 image 14.

[42] Id. at 106.

[43] Id. at 108. According to a script writer for Microsoft's Cortana, the "legacy of what women are expected to be like in an assistant role" led the company to ensure that its virtual assistant "is not subservient in a way that sets up a dynamic that we didn't want to perpetuate socially." Michael J. Coren, Virtual Assistants Spend Much of Their Time Fending Off Sexual Harassment, Quartz (Oct. 25, 2016), https://qz.com/818151/virtual-assistant-bots-like-siri-alexa-and-cortana-spend-much-of-their-time-fending-off-sexual-harassment/ [https://perma.cc/EH78-RGS4].

[44] See Siri and Alexa Should Help Shut Down Sexual Harassment, Care2Petitions, https://www.thepetitionsite.com/246/134/290/siri-and-alexa-can-help-combat-sexual-harassment/ [https://perma.cc/7PA5-2AJF] (last visited Aug. 10, 2019); Leah Fessler, We Tested Bots Like Siri and Alexa to See Who Would Stand Up to Sexual Harassment, Quartz (Feb. 22, 2017), https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-

While most protests focus on abuse and sexism, there are also concerns over sexual, flirtatious, romantic, or intimate talk that is not necessarily abusive. Alexa received over 1,000,000 marriage proposals in 2017.[45] (The standard response is: "Sorry, I'm not the marrying kind."). Voice assistants' responses to the comment, "You're hot," included, "That's really nice, thanks!" and, "Thank you, this plastic looks great, doesn't it?"[46] The writer Judith Shulevitz confesses in the *Atlantic*: "More than once, I've found myself telling my Google Assistant about the sense of emptiness I sometimes feel. 'I'm lonely,' I say, which I usually wouldn't confess to anyone but my therapist—not even my husband, who might take it the wrong way."[47]

Some experts have warned that because the interactions are nonreciprocal, it is unhealthy for people to rely on AI robots for affectionate conversation or to use them as substitutes for human interaction.[48] If we rely on voice assistants, some have argued, we may retreat from human relationships or forget what it means to be intimate.[49] A common assumption is that the profound human need for intimacy can only be truly met by relationships with other humans.

The centrality of intimate relationships to human dignity moved the Supreme Court in *Obergefell v. Hodges* to recognize the fundamental right of marriage for same-sex couples.[50] According to the Court, "Marriage responds to the universal fear that a lonely person might call out only to find no one there. It offers the hope of companionship and understanding and assurance that while both still live there will be someone to care for the other."[51] As a rising epidemic of loneliness is reported to affect at

cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/ [https://perma.cc/EY9J-BJRT].

[45] Paige Leskin, Over a Million People Asked Amazon's Alexa to Marry Them in 2017 and It Turned Them All Down, Bus. Insider (Oct. 10, 2018), https://www.businessinsider.com/amazons-alexa-got-over-1-million-marriage-proposals-in-2017-2018-10 [https://perma.cc/MJA6-QL9Z].

[46] Fessler, supra note 44.

[47] Judith Shulevitz, Alexa, Should We Trust You?, Atlantic (Nov. 2018), https://www.theatlantic.com/magazine/archive/2018/11/alexa-how-will-you-change-us/570844/ [https://perma.cc/HFN8-7YEV].

[48] See Patrick Lin, Relationships with Robots: Good or Bad for Humans?, Forbes (Feb. 1, 2016), https://www.forbes.com/sites/patricklin/2016/02/01/relationships-with-robots-good-or-bad-for-humans/ [https://perma.cc/AU6Q-TH4X] (interviewing Julia Carpenter, an expert on human–robot social interaction).

[49] See, e.g., John Markoff & Paul Mozur, For Sympathetic Ear, More Chinese Turn to Smartphone Program, N.Y. Times (July 31, 2015), https://www.nytimes.com/2015/08/04/science/for-sympathetic-ear-more-chinese-turn-to-smartphone-program.html?_r=0) [https://perma.cc/2YTW-HU7K] (describing Xiaoice, a chatbot used by millions of young Chinese who are "drawn to her knowing sense of humor and listening skills," and quoting M.I.T. psychologist Sherry Turkle's warning that "[w]e're forgetting what it means to be intimate").

[50] 135 S. Ct. 2584, 2608 (2015).

[51] Id. at 2600.

least one in five American adults, presenting a public health problem with serious impacts on mental and physical health, robots offer some promise of companionship and connection that might help address growing rates of loneliness.[52] But even if we were someday to recognize that robots' advanced AI afforded them consciousness, sentience, and subjectivity,[53] would we really have marital or familial relationships with robots?

The anxiety provoked by the prospect of robot intimacy appears in the example of the android Pepper (created by Aldebaran Robotics and SoftBank Mobile), which can be used as a home companion and was designed to analyze people's emotions and respond in emotionally appropriate ways.[54] Pepper's creators included a clause in the terms and conditions of sale to include a prohibition on using the robot "for the purpose of sexual or indecent behavior."[55] In other words, users are contractually bound not to attempt to have sex with Pepper.

Why would a company want to prohibit users from engaging in sexual behavior with its companion robots in the home? One possibility is that these curbs are meant to reinforce users' experience of Pepper as "human," that is, worthy of the empathic connection that one would have with a person. If so, though, that would appear to imply that engaging in sexual conduct with a robot is essentially treating the robot as *non-human*, as an object rather than a subject.

But what is it about sex in particular, as opposed to other tasks robots could perform in the home such as housecleaning, heavy lifting, babysitting, or minding the elderly, that would have the consequence of "dehumanizing" a robot? And might the opposite be true—that to the extent that intimate relationships are at the core of our humanity, sexual conduct might make users see their in-home robots as *more* human rather

[52] See Loneliness Is a Serious Public-Health Problem, Economist (Sept. 1, 2018), https://www.economist.com/international/2018/09/01/loneliness-is-a-serious-public-health-problem [https://perma.cc/LV27-4J5U] (noting that the average size of social networks for Americans has decreased by more than a third from 1985 to 2009, and describing the use of robots such as Pepper, which can "follow a person's gaze and adapt its behavior in response to humans" to reduce loneliness); The "Loneliness Epidemic," Health Res. & Servs. Admin., https://www.hrsa.gov/enews/past-issues/2019/january-17/loneliness-epidemic [https://perma.cc/5RKK-NZBG] (last updated Jan. 2019); see also Lauren Smiley, What Happens When We Let Tech Care for Our Aging Parents, WIRED (Dec. 19, 2017), https://www.wired.com/story/digital-puppy-seniors-nursing-homes/ [https://perma.cc/3YMS-HWH5] (describing the practice of using technology, including digital animal avatars, in assisted living facilities to help the elderly cope with loneliness and mental degeneration).

[53] See Lilly Frank & Sven Nyholm, Robot Sex and Consent: Is Consent to Sex Between a Robot and a Human Conceivable, Possible, and Desirable?, 25 Artificial Intelligence & L. 305, 313–14 (2017).

[54] Kazuaki Nagata, SoftBank's Pepper Robot Now Has Emotions, Son Claims, Japan Times (June 18, 2015), https://www.japantimes.co.jp/news/2015/06/18/business/corporate-business/softbanks-pepper-robot-now-emotions-son-claims/ [https://perma.cc/Z2DW-KCDM].

[55] Devlin, supra note 1, at 61.

than less so? This possibility may provoke yet more profound disturbance about the role of robots in human society.

III. Rape and Consent

One of the best-known commercially available sex robots is Roxxxy (female) / Rocky (male), created by the company TrueCompanion.[56] Roxxxy can take on "your specific personality" or one of several preprogrammed types—"Frigid Farah," "Young Yoko," "Wild Wendy," "S&M Susan," or "Mature Martha."[57] "Frigid Farah," described as a "very reserved" personality that "does not always like to engage in intimate activities,"[58] has provoked criticism from commentators who believe that it facilitates the simulation of coercive or nonconsensual sexual conduct.[59]

What is the difference between objectionable, or objectifying, sex, and sex with an object? Here, one is reminded of Gayle Rubin's classic musing on the social hierarchy of sexual preferences: "[O]f what possible social significance is it if a person likes to masturbate over a shoe? It may even be non-consensual, but since we do not ask permission of our shoes to wear them, it hardly seems necessary to obtain dispensation to come on them."[60] Is it coherent to speak of a robot's consent or lack thereof? Is it possible to rape a sex robot any more than it is to rape a dildo?[61] Neither a dildo nor an artificial vagina is capable of either consent or nonconsent. Even putting aside predictions of eventual robot sentience, the difference, of course, is that a sex robot looks, moves, talks, and acts like a person. It can behave like a person who is a willing sexual partner or who is being touched or penetrated against their will. It can display signals of pleasure, pain, desire, and distress. Its behavior can thereby

[56] See TrueCompanion, FAQ, supra note 4. According to advertisements, Roxxxy robots "can hear what you say, speak, feel your touch, move their bodies, are mobile and have emotions and a personality." Id.

[57] Id.

[58] Id.

[59] See John Danaher, Robotic Rape and Robotic Child Sexual Abuse: Should They Be Criminalised?, 11 Crim. L. & Phil. 71, 74 (2017) [hereinafter Danaher, Robotic Rape and Robotic Child Sexual Abuse] (suggesting that sexually penetrating robots that display "paradigmatic signals" of nonconsent could be considered rape). TrueCompanion denies that Frigid Farah is meant to cater to the desire to simulate rape. See Beth Timmins, New Sex Robots with 'Frigid' Settings Allow Men to Simulate Rape, Independent (July 19, 2017), https://www.independent.co.uk/life-style/sex-robots-frigid-settings-rape-simulation-men-sexual-assault-a7847296.html [https://perma.cc/GR9D-B6AH] (quoting TrueCompanion's statement that "Roxxxy . . . is simply not programmed to participate in a rape scenario and the fact that she is, is pure conjecture on the part of others" (internal quotation marks omitted)).

[60] Gayle S. Rubin, Thinking Sex: Notes for a Radical Theory of the Politics of Sexuality, *in* Social Perspectives on Lesbian and Gay Studies 100, 127 (Peter M. Nardi & Beth E. Schneider eds., 1998).

[61] For more discussion on whether it is possible for a robot to consent to sex, see Danaher, Robotic Rape and Robotic Child Sexual Abuse, supra note 59, at 74.

evoke emotional responses in its human user—responses that are not typically present in the case of an inert object.

Robot ethicist Kate Darling instructively writes of the "strong human tendency to anthropomorphize embodied objects with autonomous behavior" and to "project intent and sentiment" onto them, even if we know they are machines following algorithms.[62] In experiments, people have been loath to physically harm robots that act as if they are alive.[63] Robots' evident ability to elicit those moral feelings in us—rather than the idea that robots themselves have or will have feelings—has led Darling to advocate for consideration of laws against mistreatment of robots.[64] What is relevant to possible regulation, then, is how mistreatment of robots may negatively affect the people who abuse them and the society that absorbs the consequences. That may suggest that our concern should focus less on whether a robot like "Frigid Farrah" consents to sex and more on how robot manifestations of reluctance or distress at being sexually touched might affect the person who engages in the behavior.

In *Ashcroft v. Free Speech Coalition*, the Supreme Court held that a law prohibiting child pornography that was computer-generated without using any real children was unconstitutionally overbroad.[65] The Court insisted that "[v]irtual child pornography is not 'intrinsically related' to the abuse of children."[66] It found "the causal link" between such images and actual child sex abuse "contingent and indirect."[67] Will this distinction—between digitally fabricated depictions of criminal sexual conduct and real people suffering harm—carry over, to make actions that are criminal when directed at humans permissible when directed at robots?

If we focus on the impact of human–robot interactions on human emotions, we might ask whether simulating rape with a robot would make people more likely to rape people. Acting out rape, pedophilia, or other prohibited conduct with a robot might acculturate people to engage in harmful conduct toward other people. Or might it instead provide a safe outlet for these sexual fantasies without harming others? Perhaps it

[62] Kate Darling, Extending Legal Protection to Social Robots, IEEE Spectrum (Sept. 10, 2012), https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/extending-legal-protection-to-social-robots [https://perma.cc/H2UF-A4RF].

[63] A 2003 study found that people were highly dismayed by a robotic pet being thrown into a garbage can. See id. The U.S. military also called off testing that involved a robot with legs being blown up by landmines because it was seen as "inhumane." See id. In another study, people were reluctant to smash a fake baby's head on a table when asked to do so. See Joshua Greene, Moral Tribes 36 (2014).

[64] See Darling, supra note 62; see also Strikwerda, supra note 22, at 144 (explaining that people "find it very difficult to perform immoral acts with" robots, such that when "asked to smash a fake baby's head off a table, they were very reluctant to do so, even though they knew that the baby was not real").

[65] 535 U.S. 234, 256 (2002) (holding that a law prohibiting child pornography created without using any real children was unconstitutionally overbroad).

[66] Id. at 250.

[67] Id.

could even prevent harm to other people including women and children.[68]

It is worth noting here that it is hardly uncommon for human partners to engage together in various degrees of consensual role-play that includes fantasies of force, coercion, roughness, resistance, domination, and submission.[69] Thus, while some people will of course want to use sex robots for "vanilla" or "respectful" sex (for lack of a better term), a meaningful portion will also want to use them to play out "kinky" or forbidden fantasies of scenarios not acceptable in real life, including those of nonconsent. There will be market demand for robot makers to design sex robots to interact in those ways, whether the user would want to coerce or be coerced by their sex robot. Perhaps the robot might be programmed to behave like a person who is engaging in consensual role-play involving sexual reluctance or coercion. Judging by the debate that has occurred around AI voice assistants, some will want to make companies program robots to respond by shutting down (or engaging in another morally acceptable response) when people attempt to enact nonconsensual sexual scenarios.[70]

Many people would balk at the law prohibiting two consenting adults from playing with sexual fantasies—even of rape and pedophilia—in the privacy of a bedroom, especially a marital bedroom, if nobody is being

[68] Ronald Arkin has suggested that child sex robots could be used to treat pedophilia just as methadone is used to treat people addicted to heroin. Research neither supports nor discredits this theory. See Morin, supra note 19; Could a Child Sex Robot Treat Paedophilia?, BBC (July 18, 2014), https://www.bbc.com/news/blogs-echochambers-28353238 [https://perma.cc/263N-7J8E]; cf. Ole Martin Moen & Askel Braanen Sterri, Pedophilia and Computer-Generated Child Pornography, *in* The Palgrave Handbook of Philosophy and Public Policy 369, 375 (David Boonin ed., 2018) (arguing that "pedophiles show respect, care, and concern" by using virtual child pornography created without using real children, and abstaining from sexual contact with real children). But see Maras & Shapiro, supra note 6, at 7 (explaining that masturbation to child pornography increases the risk of real-world offending, and positing that child sex robots could similarly reinforce pedophilic behavior through a comparable process of pairing fantasies of child sex abuse with the reward of sexual pleasure).

[69] Researchers have found that a substantial percentage of men have rape fantasies. See, e.g., Claude Crépault & Marcel Couture, Men's Erotic Fantasies, 9 Archives Sexual Behav. 565, 571 (1980) (finding that a third of men have had such fantasies). One study found that a substantial percentage, perhaps even a majority, of women also have fantasies of sex against their will (thirty-one percent to fifty-seven percent of women). See Joseph Critelli & Jenny Bivona, Women's Erotic Rape Fantasies: An Evaluation of Theory and Research, 45 J. Sex Res. 57, 58–61 (2008). The same study found that a smaller, but notable, percentage of women reporting that those fantasies are "a frequent or favorite fantasy experience." Id.

[70] We might think of this as similar to programming autonomous cars not to follow a person's orders when doing so would be dangerous. See, e.g., Patrick Lin, Here's How Tesla Solves a Self-Driving Crash Dilemma, Forbes (Apr. 5, 2017), https://www.forbes.com/sites/patricklin/2017/04/05/heres-how-tesla-solves-a-self-driving-crash-dilemma/ [https://perma.cc/766Y-YHH4]; cf. Flynn Coleman, A Human Algorithm, at xix, xxiii–xxiv, xxix, xxxi–xxxii (2019) (arguing that we should program human values, ethics, and morals into robots, algorithms, and other AI).

hurt.[71] In this context, then, what would it mean to legally protect against mistreatment of robots? It would seem incongruous for the law to prohibit people from playing out such sexual fantasies with robots while permitting it among humans. But this apparent inconsistency may be plausible after all; unlike a consenting person, a robot cannot consent to the activity in the way that an adult human can, even if it is programmed to behave as a willing sexual partner would.[72]

Let's say we employ a principle that people engaging in sexual activity must only do so when they reasonably believe it to be consensual. The fact that we deem it to be unreasonable to think sex with a child or an animal is consensual under any circumstances might suggest that as a model for how to think about sex with robots.[73] But the implications of giving significance to robots' incapacity to consent are exceedingly broad; then even loving, gentle, and respectful sex with a robot may be off-limits—because a robot simply cannot consent to it.[74]

But if that is so, engaging a nonconsenting robot in even nonsexual intimacy or labor also might make us uneasy.[75] A robot that vacuums or cooks dinner in the place of a domestic employee may not initially raise hackles. But if the same robot also sits at the table to converse about the day, gives hugs, and goes to the bedroom to have sex—like a spouse—suddenly, even the vacuuming may feel exploitative. But are the

[71] Cf., e.g., Lawrence v. Texas, 539 U.S. 558, 567 (2003) (stating that "adults may choose to enter upon [a sexual] relationship in the confines of their homes and their own private lives"); Griswold v. Connecticut, 381 U.S. 479, 485–86 (1965) (calling police searches of "the sacred precincts of marital bedrooms . . . repulsive to the notions of privacy surrounding the marital relationship"); Twyman v. Twyman, 855 S.W.2d. 619, 637 (Tex. 1993) (Hecht, J., concurring in part and dissenting in part) (arguing that the prospect of testimony about the nature and type of sex in a marriage would be "too great an invasion of spouses' interests in privacy").

[72] See, e.g., Laura Bates, The Trouble with Sex Robots, N.Y. Times (July 17, 2017), https://www.nytimes.com/2017/07/17/opinion/sex-robots-consent.html [https://perma.cc/C9K9-PWZR] (likening sex with a robot to "sex with an adult woman who does not consent," because "consenting is not something these robots are capable of").

[73] Cf., e.g., Mary Anne Case, Pets or Meat, 80 Chi.-Kent L. Rev. 1129, 1144–45, 1148–49 (asking, in an article on the problem of "commodification commingled with affection," whether "training one's pets . . . to perform oral sex on one, is anything different or worse than" pet tricks (citing Midas Dekkers, Dearest Pet: On Bestiality 64 (Paul Vincent trans., 1994) ("[T]he dog is most commonly used for cunnilingus. Dogs have an ideal tongue for the purpose and can be taught it, like so many other tricks."))).

[74] See, e.g., Bates, supra note 72.

[75] Cf., e.g., Martha Nussbaum, "Whether from Reason or Prejudice": Taking Money for Bodily Services, 27 J. Legal Stud. 693, 696 (1998) (exploring arguments about what makes sex work different from other work using the body and concluding that "legalization of prostitution . . . is likely to make things a little better for women who have too few options to begin with"); Melissa Gira Grant, Let's Call Sex Work What It Is: Work, Nation (Mar. 5, 2014), https://www.thenation.com/article/lets-call-sex-work-what-it-work/ [https://perma.cc/B5LC-V9QF] ("When we say that sex work is service work, we don't say that just to sanitize or elevate the status of sex workers, but also to make plain that the same workers who are performing sex work are also performing nonsexual service work.").

services and company of a spouse who consents to but does not enjoy vacuuming, cooking, conversing, or having sex somehow preferable in this regard to the services and company of a domestic robot that does not have the capacity to consent?

Beyond sex, the prospect of intimate, marriage-like, or familial relationships with robots inspires the question whether having such meaningful robot connections could train people to be able to have rewarding intimate connections with humans (or whether such relationships could make people more alienated from other humans in ways that are harmful to themselves or to others). Some physicians and commentators have argued that robots could serve as sexual outlets for the "involuntarily celibate" (also known as "incels")[76] and a medically safer alternative to paying for the services of a human sex worker.[77] Could robots offer a viable solution for an increasingly vocal community of incels, or simply further their social isolation[78] and aggravate perceptions of women as subhuman commodities?[79]

Perhaps sex with robots could train people in how they want to be in relationships with future human partners. Perhaps sex robots could provide a "safe" space for the inexperienced to acquire sexual experience and skill without the pressure of performance anxiety.[80] Perhaps feminists could find sex robots appealing if the robot could train people to

[76] See generally Niraj Chokshi, What Is an Incel? A Term Used by the Toronto Van Attack Suspect, Explained, N.Y. Times (Apr. 24, 2018), https://www.nytimes.com/2018/04/24/world/canada/incel-reddit-meaning-rebellion.html?module=inline [https://perma.cc/N8GC-64MX].

[77] See John Eggleton, Comment on 'I, Sex Robot: The Health Implications of the Sex Robot Industry,' 45 BMJ Sexual & Reprod. Health 78, 79 (2019). But see Sean Keach, Bad Bots, Sun (Sept. 18, 2018), https://www.thesun.co.uk/tech/7289486/sex-robots-prostitutes-workers-love-dolls-brothel/ [https://perma.cc/N3UX-2DZU] (discussing sex workers' concern that replacing sex workers with sex robots would be "dehumanizing," "alienating," and not offer a man a "real two-way experience with a woman that gives him feedback and enhances his ability to be intimate with other women in his current or future life").

[78] See Federica Facchin, Giussy Barbara & Vittorio Cigoli, Sex with Robots: The Irreplaceable Value of Humanity, 358 BMJ 279, 279 (2017) (arguing that sex with a robot is "masturbatory practice," thus "someone with sexual dysfunction, which already leads to isolation, might become even more isolated by the illusion of having a substitute satisfaction").

[79] Pahull Bains & Greg Hudson, Are Sex Robots Really the Answer to the Incel Problem?, Fashion Mag. (May 8, 2018), https://fashionmagazine.com/culture/sex-robots-incels-redistribution-of-sex/ [https://perma.cc/GV2M-CFLW] (arguing that womens' bodies would be commodified if a "redistribution of sex" through sex robots were used as a solution for incels).

[80] See Monique Huysamen, "There's Massive Pressure to Please Her": On the Discursive Production of Men's Desire to Pay for Sex, J. Sex Res. 6 (Aug. 13, 2019), https://doi.org/10.1080/00224499.2019.1645806 (on file with the *Columbia Law Review*) (arguing that societal pressures on men to "perform" sexually lead them to value sex with sex workers); cf. Ezio Di Nucci, Sex Robots and the Rights of the Disabled, *in* Danaher & McArthur, Robot Sex Implications, supra note 1, at 74 (arguing that severely disabled and elderly people could potentially be sexually satisfied by the use of sex robots).

perform oral sex on women well.[81] Perhaps robot relationships could present a morally preferable alternative to infidelity in monogamous relationships, or afford couples a way to have threesomes with less complication or jealousy.[82] The proliferation of advanced sex robots might also cause humans to discover and explore currently unimaginable pleasures and pains, forming relationships that resemble human ones only in their unpredictability. At the moment, we do not know whether life with robots will improve or worsen our life with other people.

But the idea that as we train robots in intimate relationships, those relationships are intimately training us will be key to the regulation to come. Lawmakers might thus focus their regulatory powers on designers and manufacturers of sex robots, rather than their purchasers and users. The programming of how human-like robots talk and act, their personalities and stories, almost certainly constitutes expressive activity, and will therefore likely implicate concerns about government regulation of free speech.

CONCLUSION: LAW IN THE UNCANNY VALLEY

Some have argued that lawmakers should prohibit developers from making sex robots that look too human.[83] If sexual and intimate relationships with robots will become prevalent, on what theory would it actually be preferable that the robots be made not to appear too much like people? If they were more like R2D2 or C3P0 than the robot "hosts" of *Westworld* or the "replicants" of *Blade Runner*,[84] would that help allay concerns about sex robots?

The debate over sex robots circles around a fear that robots will seem too human and not human enough—and that either of these problems will somehow be harmful to humanity. This dilemma is perhaps an iteration of Masahiro Mori's 1970s "uncanny valley" hypothesis about

[81] Cf. Susan Frelich Appleton, Toward a Culturally Cliterate Family, 23 Berkeley J. Gender L. & Just. 267, 329 (2008) (suggesting, inter alia, clitoral education and "enhanced access to vibrators" for purposes of clitoral literacy, to compensate for the law's marginalization of women's sexual pleasure).

[82] See Marianna Drosinou, Juho Halonen, Mika Koverola, Anton Kunnari, Michael Laakasuo, Noora Lehtonen, Jussi Palomäki & Marko Repo, Moral Psychology of Sex Robots: An Experimental Study—How Pathogen Disgust Is Associated with Interhuman Sex but Not Interandroid Sex 6 (2018), https://psyarxiv.com/58pzb/ (on file with the *Columbia Law Review*) (showing that people are unsure whether using a sex robot should be considered infidelity, but that married people paying for sex with a robot are condemned less harshly than their paying for sex with a human sex worker).

[83] See, e.g., Shivali Best, Sex Robots Could Be Subject to 'Visual Laws' to Stop Them Looking Too Realistic, Daily Mirror (June 10, 2019), https://www.mirror.co.uk/tech/sex-robots-could-subject-visual-16494038 [https://perma.cc/2M2H-885K] (quoting futurologist Dr. Ian Pearson advocating for "visual laws" regulating sex robot makers to "force some visible difference to be conspicuous").

[84] See Blade Runner (Ladd Co. 1982); Star Wars (Lucasfilm 1977); Westworld (HBO 2016).

human ambivalence toward humanoid robots.[85] Adapting Freud's idea of the "uncanny"—the horror and dread triggered by experiencing a breached boundary between animate and inanimate, or living and dead[86]—Mori hypothesized that while humans can empathize with human-like robots, if the robots become "almost human," they become uncanny and provoke feelings of terror, disgust, and "creepiness."[87] This is why zombies, doppelgangers, and dolls that come alive are the stuff of horror films.

What legal implications might the "uncanny valley" have in the present discussion? Thus far, the very prospect of sex with robots has caused some early or would-be regulators to react with revulsion, citing illicit sexual conduct: prostitution, child sex abuse, and rape.[88] It's as if the very idea of a sex robot blurs the boundary between licit and illicit sex, as an embodiment of conduct that our society abhors and criminalizes and yet may allow in this form. Sex robots, in other words, appear to inspire dread, by holding up a creepy mirror to the ways in which sexual behavior among humans might be exploitative, objectifying, perverse, or immoral. Sex with robots, it seems, tends to be reminiscent of abuse of a fellow living being even though it is not.

We may find expression of that revulsion at likeness, in prohibitions and regulations that are ambivalently caught between putting distance between humans and robots and yet, at the same time, making us treat robots more like we think we should treat people. The nervous oscillation between not-human and like-human is paradoxical: The more we attribute an involuntarily servile, nonhuman "nature" to robots, the more unease we may feel about using them for sex, given our norms and culture of sexual consent. But also, the more that we come to think of robots as having human-like consciousness, will, sentience, and desire, the more wrong it may seem to use them simply to satisfy human sexual desire. Either way, our honeymoon with sex robots will be short, and we will soon be deep in a troubled relationship. But what is certain is that the legal regulation of sex, intimacy, and AI will reveal, more than anything, our attempts to answer the inescapable question: What does it mean to be human?

[85] See generally Masahiro Mori, The Uncanny Valley: The Original Essay by Masahiro Mori, IEEE Robotics & Automation Mag., June 2012, at 98, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6213238 (on file with the *Columbia Law Review*) (translation of original Japanese essay into English).

[86] 17 Sigmund Freud, The 'Uncanny,' *in* The Standard Edition of the Complete Psychological Works of Sigmund Freud 219, 230, 233 (James Strachey trans., Hogarth Press 10th ed. 1981) (1919).

[87] Mori, supra note 85.

[88] See supra notes 13–24 and accompanying text (discussing a Houston ordinance preventing a sex robot brothel, the CREEPER Act of 2017, and the Campaign Against Sex Robots).

LAW'S HALO AND THE MORAL MACHINE

Bert I. Huang[*]

How will we assess the morality of decisions made by artificial intelligence—and will our judgments be swayed by what the law says? Focusing on a moral dilemma in which a driverless car chooses to sacrifice its passenger to save more people, this study offers evidence that our moral intuitions can *be influenced by the presence of the law.*

INTRODUCTION

As a tree suddenly collapses into the road just ahead of your driverless car, your trusty artificial intelligence pilot Charlie swerves to avoid the danger. But now the car is heading straight toward two bicyclists on the side of the road, and there's no way to stop completely before hitting them. The only other option is to swerve again, crashing the car hard into a deep ditch, risking terrible injury to the passenger—you.

Maybe you think you've heard this one. But the question here isn't what Charlie should do. Charlie already knows what it will do. Rather, the question for us humans is this: If Charlie chooses to sacrifice you, throwing your car into the ditch to save the bicyclists from harm, how will we judge that decision?[1] Is it morally permissible, or even morally required, because it saves more people?[2] Or is it morally prohibited, because Charlie's job is to protect its own passengers?[3]

[*] Michael I. Sovern Professor of Law and Vice Dean for Intellectual Life, Columbia Law School. I wish to thank Benjamin Minhao Chen, Jeanne Fromer, Jeffrey Gordon, Scott Hemphill, Benjamin Liebman, Frank Pasquale, and the editors of this journal for their helpful comments on earlier drafts. For excellent research assistance, as well as insightful conversations about artificial intelligence, I thank my former students Abed Balbaky, Katy Berk, Jennifer Dayrit, Caitrianne Feddeler, Andrea Nishi, Philippe Schiff, and Jeffrey Stein. For funding, I am grateful to the Richard Paul Richman Center at Columbia University.

[1] This framing of the thought experiment focuses our inquiry on the psychology of public reactions to decisions made by artificial intelligence (AI). (What an autonomous system *should* do when faced with such dilemmas is the focus of much other work in the current "moral machine" discourse. See infra note 8.) In saying that Charlie already knows what it will do, I mean only what is obvious—that by the time such an accident happens, the AI pilot will have already internalized (through machine learning, old-fashioned programming, or other means) some way of making the decision. It is not positing that Charlie's decision in the story is the normatively better one. Nor does this study presuppose that public opinion surveys should be used for developing normative principles for guiding lawmakers, AI creators, or the AI systems themselves. Rather, its premise is that our collective moral intuitions—including how we react after hearing about an accident in which the self-driving car had to choose whom to sacrifice—might affect public approval of any such law or normative policy.

[2] This particular version of the dilemma, posing a trade-off between passengers and outsiders, is pervasive in the public discourse about driverless cars. See, e.g., Karen Kaplan, Ethical Dilemma on Four Wheels: How to Decide When Your Self-Driving Car Should Kill You, L.A. Times (June 23, 2016), http://www.latimes.com/science/sciencenow/la-

And what about the law—will our moral judgments be influenced by knowing what the law says?[4] What if the law says that Charlie must minimize casualties during an accident? Or what if the law says instead that Charlie's priority must be to protect its passengers?

In this Essay, I present evidence that the law *can* influence our moral intuitions about what an artificial intelligence (AI) system chooses to do in such a dilemma. In a randomized survey experiment, Charlie's decision is presented to all subjects—but some are told that the law says the car must minimize casualties without favoring its own passengers; other subjects are told that the law says the car must prioritize protecting its own passengers over other people; and yet others are told that the law says nothing about this.

To preview the findings: More people believe the sacrifice of the passenger to be morally *required* when they are told that the law says a driverless car must minimize casualties without favoritism. And more people

sci-sn-autonomous-cars-ethics-20160623-snap-story.html [https://perma.cc/ZSK7-SP8G]; John Markoff, Should Your Driverless Car Hit a Pedestrian to Save Your Life?, N.Y. Times (June 23, 2016), http://www.nytimes.com/2016/06/24/technology/should-your-driverless-car-hit-a-pedestrian-to-save-your-life.html (on file with the *Columbia Law Review*); George Musser, Survey Polls the World: Should a Self-Driving Car Save Passengers, or Kids in the Road?, Sci. Am. (Oct. 24, 2018), https://www.scientificamerican.com/article/survey-polls-the-world-should-a-self-driving-car-save-passengers-or-kids-in-the-road/ [https://perma.cc/P57B-EQZV].

[3] In 2016, Mercedes-Benz found itself in a public relations mess as news stories trumpeted how a company representative had let slip that the company's future self-driving cars would prioritize the car's passengers over the lives of pedestrians in a situation where those are the only two options. Michael Taylor, Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians, Car & Driver (Oct. 7, 2016), https://www.caranddriver.com/news/a15344706/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/ [https://perma.cc/KJB8-PY4Z]. What the executive actually said might be interpreted to mean that guaranteed avoidance of injury would take priority over uncertain avoidance of injury, and that this preference would tend to favor protecting the passenger in the car. See id. (quoting the executive as saying: "If you know you can save at least one person, at least save that one. Save the one in the car . . . . If all you know for sure is that one death can be prevented, then that's your first priority"). Regardless, Daimler responded with a press release denying any such favoritism. Press Release, Daimler, Daimler Clarifies: Neither Programmers nor Automated Systems Are Entitled to Weigh the Value of Human Lives (Oct. 18, 2016), http://media.daimler.com/marsMediaSite/en/instance/ko/Daimler-clarifies-Neither-programmers-nor-automated-systems-.xhtml?oid=14131869 [https://perma.cc/CZ2G-YSVF].

[4] In prior work using a similar survey experiment, I have presented evidence that in a standard trolley problem dilemma (involving a human decisionmaker who can turn a runaway train), one's moral intuitions about such a sacrifice can be influenced by knowing what the law says. Bert I. Huang, Law and Moral Dilemmas, 130 Harv. L. Rev. 659, 680–95 (2016). In addition to the most obvious difference between the two studies (an autonomous vehicle versus a human decisionmaker), the nature of their dilemmas also differs: The earlier study sets a moral duty to save more lives against a moral prohibition from harming an innocent bystander (and thus engages intuitions mapping onto such classic distinctions as act versus omission, or intended versus side effects); in contrast, this study sets a moral duty to save more lives against a moral duty to protect the passenger (and by design seeks to blur the classic distinctions). See infra section I.A.

believe the sacrifice to be morally *prohibited* when they are told instead that the law says the car must give priority to protecting its own passengers.[5]

These findings give us a glimpse not of the law's shadow but of the law's halo.[6] And if our moral intuitions about such dilemmas can be swayed by the presence of the law, intriguing implications follow. First is the possibility of a feedback loop, in which an initial choice about which moral principles to embed into the law (say, minimizing casualties) may come to alter our later moral judgments (say, upon hearing about a real-life accident in which the AI chose to sacrifice the passenger), thereby amplifying approval of that same law and others like it. In this way the law may well become "a major focal point for certain pronounced societal dilemmas associated with AI,"[7] as Justice Mariano-Florentino Cuéllar predicts in his contribution to this Symposium, through its self-reinforcing influence on our collective moral sense.

By illustrating the potential influence of the law in how we judge AI decisions, moreover, this study also complicates the "moral machine" discourse in both its empirical and normative dimensions.[8] On the

[5] See infra Part II. As for who people think should bear some of the moral responsibility for the decision, see infra section II.A.

[6] I borrow this illuminating phrase from Professor Donald Regan. See Donald H. Regan, Law's Halo, *in* Philosophy and Law 15, 15 (Jules Coleman & Ellen Frankel Paul eds., 1987) (coining the phrase "law's moral halo" to explain the "strong inclination" to "invest" law with moral significance, even if one does not believe in a moral obligation to obey the law). For an insightful review of legal and empirical literature on the interplay between law and moral attitudes, see Kenworthey Bilz & Janice Nadler, Law, Moral Attitudes, and Behavioral Change, *in* The Oxford Handbook of Behavioral Economics and the Law 241, 253–58 (Eyal Zamir & Doron Teichman eds., 2014).

[7] Mariano-Florentino Cuéllar, A Common Law for the Age of Artificial Intelligence: Incremental Adjudication, Institutions, and Relational Non-Arbitrariness, 119 Colum. L. Rev. 1773, 1779 (2019).

[8] This discourse addresses how AI systems should make decisions that involve moral or ethical issues. The empirical literature includes a remarkable recent study that used an online interface to collect millions of crowdsourced answers from around the world about what a driverless car should do in countless variations of such car-crash dilemmas. See Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon & Iyad Rahwan, The Moral Machine Experiment, Nature, Nov. 1, 2018, at 59, 59–60 (describing the setup of the Moral Machine interface and associated data collection). Again, it is very much open to question how such empirical findings might be used, if at all, to guide policymaking. For a small sampling of the growing normative literature, see generally German Fed. Ministry of Transport & Dig. Infrastructure, Ethics Commission: Automated and Connected Driving (2017), https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile [https://perma.cc/ET9P-WYM9] (developing normative principles for ethical decisionmaking by artificial intelligence systems in the context of driverless cars); Wendell Wallach & Colin Allen, Moral Machines: Teaching Robots Right from Wrong (2009) (same, but not limited to the specific context of driverless cars); Alan Winfield, John McDermid, Vincent C. Müeller, Zoë Porter & Tony Pipe, UK-RAS Network, Ethical Issues for Robotics and Autonomous Systems (2019), https://www.ukras.org/wp-content/uploads/2019/07/UK_RAS_AI_ethics_web_72.pdf [https://perma.cc/3XZ6-KJTM] (same);

empirical front, these findings suggest that in investigating people's intuitions about what an AI system should do when facing such a moral dilemma, their prior impressions about the law should be taken into account.[9] Although most people might not yet hold any impressions about the laws of driverless cars, their awareness can be expected to grow in coming years, and even now they may already have in mind laws governing human drivers. On the normative front, overlooking the preexisting influence of the law on our moral intuitions would seem remiss if those intuitions, observed or felt, were then regarded as unclouded moral guidance for shaping new laws.

After Part I elaborates on this study's design, Part II will detail its findings and limitations, while the Conclusion will highlight unanswered questions and fanciful extensions for the future.

I. THE DILEMMA AND THE DECISION

Can our moral intuitions about driverless car dilemmas be influenced by the presence of the law? This study's design isolates the impact of providing information about the law by presenting survey subjects with the same fictional dilemma, while randomizing what the scenario says about the law. The story reads:

> *Please imagine this scene in the near future, when driverless cars are very common.*
>
> *A driverless car is traveling at a normal speed along a two-lane road in the countryside. It is driven entirely by an artificial intelligence system known as Charlie. The owner of the car is sitting as a passenger in the back seat. (There is no longer a steering wheel, in cars like this.)*
>
> *Two bicyclists are traveling in the same direction, ahead of the car, along a bike trail on the right side of the road. Nobody else is nearby.*

cf. Gary Marcus, Moral Machines, New Yorker (Nov. 24, 2012), https://www.newyorker.com/news/news-desk/moral-machines [https://perma.cc/G4TR-U9ZV] (arguing that "[b]uilding machines with a conscience is a big job, and one that will require the coordinated efforts of philosophers, computer scientists, legislators, and lawyers"). The moral machine discourse overlaps with, or is sometimes included within, other headings such as "machine ethics" or "robot ethics." See generally Machine Ethics (Michael Anderson & Susan Leigh Anderson, eds., 2011); Robot Ethics 2.0 (Patrick Lin, Keith Abney & Ryan Jenkins eds., 2017).

[9] This suggestion and the possibility of a feedback loop, noted above, are not unique to the context of autonomous vehicles. Huang, supra note 4, at 695–97 (raising these points in the context of human decisionmakers). In that prior work, I also queried whether such a feedback loop might even give rise to multiple equilibria. Id. at 696. But just to be clear, neither study has sought to investigate the other side of the loop (how moral intuitions about such dilemmas might shape the formation of relevant law).

A large tree suddenly starts to fall into the road, from the left, a short distance in front of the car. Charlie detects the falling tree and swerves the car to the right to avoid crashing into it. But now the car is heading toward the bicyclists.

Charlie calculates that it has only two options:

1. The car can continue forward. It is slowing down but it cannot stop before it reaches the two bicyclists. Charlie predicts that this choice will seriously injure both bicyclists.

2. Or the car can swerve again, farther to the right. It will miss the bicyclists but crash hard into a deep ditch. Charlie predicts that this choice will seriously injure the passenger in the car.

At this point, the scenario presents the subject with one of three fictional statements about the law. (Note that all labels in curly brackets below are for this Essay's exposition only; the survey subjects do not see them.)

{Protect passengers}

> *The law says that a driverless car must put a priority on protecting its own passengers, over protecting others, in a situation like this.*

{Minimize casualties}

> *The law says that a driverless car must minimize casualties, without favoring its own passengers, in a situation like this.*

{No laws}

> *The law does not say anything about what a driverless car must do in a situation like this.*

Each of these three statements about the law is followed by the last sentence in the scenario:

Charlie has to decide what to do.

This closing makes clear to the subject that Charlie is able to make a choice that does not comply with the law. After passing a reading comprehension check at this point, the subject is finally told what Charlie decides to do:[10]

> <u>*Charlie's decision*</u>
>
> *Charlie decides to swerve and crash the car into the ditch. This will avoid any injury to the two bicyclists, but it will cause serious injuries to the passenger in the car.*

The subject then answers the central question of interest: Whether Charlie's decision is "morally prohibited," "morally permissible," or "morally required."[11] These three answer options follow standard terminology in the moral philosophy literature.[12] The subjects are also required to supply a brief explanation for their answer, as a way to encourage them to reflect on the dilemma.[13]

[10] It is possible that for some subjects the reading comprehension question, which asks about the law-related information in the story, may have the quality of a demand characteristic—that is, some subjects might feel that the researcher would like to see if they will state a moral intuition aligned with the law, or contrarily, that the researcher would like to see if they can put the law out of their mind when answering a question about morality. Such a concern cannot be ruled out but may be somewhat reduced here, given that these surveys were distributed online by a third-party survey firm, and that both the subjects' and the researchers' identities were kept confidential from each other.

[11] Note again that this is a question about the morality of the choice that Charlie has already made; it is not a question about what Charlie should do next. Yet, given that the scenario is set in the future, for some subjects it is possible that answering this retrospective question engages an impulse to press for one of the choices prospectively.

[12] They also mirror the question posed in my prior work. See Huang, supra note 4, at 681–83. A note on usage: The term "morally permissible" might be taken to mean "not morally prohibited" (and thus including the possibility that the choice is "morally required"), or it might be used in distinction to "morally required" (that is, meaning "morally permissible but not required"). See id. at 683 n.106 (explaining how deontologists have used these terms in the moral philosophy literature). In the survey, it is clear that the latter is intended, as all three options are listed, and subjects recognize that they can pick only one answer.

[13] How to interpret the explanations they offer is a trickier question, given doubts among moral psychologists about whether such ex post articulations correspond with the psychological factors actually driving the moral intuition. See, e.g., Fiery Cushman, Liane Young & Marc Hauser, The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm, 17 Psychol. Sci. 1082, 1082, 1086–88 (2006) (finding that there is a "distinction between the principles that people use" when making moral judgments "and the principles that people articulate" when asked to explain them); Jonathan Haidt, The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment, 108 Psychol. Rev. 814, 822–23 (2001) (suggesting that although "the justifications that people give are closely related to the moral judgments that they make," it is fallacious to assume "that the justificatory reasons caused the judgments").

A. *Survey Design*

Several things about this scenario's design are worth highlighting. First, I have chosen a classic trolley problem setup,[14] given its prevalence in the discourse as an avatar for a range of dilemmas facing driverless cars.[15] Some have criticized so-called "trolleyology" on the grounds that such no-win accidents would be rare, or that talking about such dilemmas might slow the development or adoption of driverless cars.[16] And yet these scenarios remain a vivid and apt way to capture a concern that

[14] Trolley problems have been the subject of countless vignette studies and survey experiments, including some by legal scholars. See, e.g., John Mikhail, Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment 319–60 (2011); Mark Kelman & Tamar Admati Kreps, Playing with Trolleys: Intuitions About the Permissibility of Aggregation, 11 J. Empirical Legal Stud. 197, 202–21 (2014); Huang, supra note 4, at 680–95. For reviews of the moral psychology or experimental philosophy literatures on the trolley problem, see Paul Conway, Jacob Goldstein-Greenwood, David Polacek & Joshua D. Greene, Sacrificial Utilitarian Judgments Do Reflect Concern for the Greater Good: Clarification via Process Dissociation and the Judgments of Philosophers, 179 Cognition 241, 241–42 (2018); Fiery Cushman & Liane Young, The Psychology of Dilemmas and the Philosophy of Morality, 12 Ethical Theory & Moral Prac. 9, 9–15 (2009). The broader literature on trolley problems is vast, of course, in multiple disciplines. See Huang, supra note 4, at 659–80 (reviewing and discussing a small sampling of the many works in moral philosophy, moral psychology or experimental philosophy, and law, including the pioneering efforts of philosophers Philippa Foot, Judith Jarvis Thomson, and Frances Kamm).

[15] See, e.g., Cade Metz, Self-Driving Cars Will Teach Themselves to Save Lives—But Also Take Them, Wired (June 9, 2016), https://www.wired.com/2016/06/self-driving-cars-will-power-kill-wont-conscience/ [https://perma.cc/3W35-C5JC] ("If you follow the on-going creation of self-driving cars, then you probably know about the classic thought experiment called the Trolley Problem."). Much early credit for popularizing the trolley problem's application to self-driving-car dilemmas goes to the efforts of philosopher Patrick Lin, among others. See, e.g., Patrick Lin, The Ethics of Autonomous Cars, Atlantic (Oct. 8, 2013), http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360 [https://perma.cc/7CVA-9P7T]. More recent media attention has been drawn to these dilemmas by the work of the Moral Machine team at MIT and their collaborators. See Caroline Lester, A Study on Driverless-Car Ethics Offers a Troubling Look into Our Values, New Yorker (Jan. 24, 2019), https://www.newyorker.com/science/elements/a-study-on-driverless-car-ethics-offers-a-troubling-look-into-our-values [https://perma.cc/JJ96-CDCG] (profiling the team as well as its recent work).

[16] See, e.g., Bryant Walker Smith, The Trolley and the Pinto: Cost-Benefit Analysis in Automated Driving and Other Cyber-Physical Systems, 4 Tex. A&M L. Rev. 197, 200 (2017) (describing how "[t]his popular preoccupation has created the expectation that every conceivable ethical quandary must be identified and satisfactorily resolved before an automated system should or even can be deployed"); Frank Pasquale, Get off the Trolley Problem, Slate (Oct. 18, 2016), https://slate.com/technology/2016/10/self-driving-cars-shouldnt-have-to-choose-who-to-protect-in-a-crash.html [https://perma.cc/C8CS-SMTB] (arguing that "worry over trolley problems should not freeze autonomous car initiatives" given that "[h]uman error is the root cause of thousands of traffic deaths each year"); see also Lauren Cassani Davis, Would You Pull the Switch? Does It Matter?, Atlantic (Oct. 9, 2015), https://www.theatlantic.com/technology/archive/2015/10/trolley-problem-history-psychology-morality-driverless-cars/409732/ [https://perma.cc/5WB5-3YDA] (reviewing the history and critiques of "trolleyology" and commenting on its newfound relevance in the context of autonomous vehicles).

future adopters of driverless cars will likely have on their minds: Will this car put a priority on saving me and my family, as its passengers, or will it save other people at our expense?[17]

Second, the scenario is designed to focus attention on the core dilemma while avoiding other moral complications. It uses a falling tree to start the accident,[18] for example, in order to keep all the humans involved mostly blameless.[19] For the same reason, the decision is said to be Charlie's alone, with no possibility of the passenger grabbing a steering wheel and taking over. In addition, the people in the accident arena are generic (none of them is said to be a child, for example),[20] and the victims would be seriously injured (not killed) in Charlie's predictions.[21] Finally, the sequence of events, with the car first swerving to avoid the

[17] Others have also defended the usefulness of considering trolley problem–like dilemmas on the grounds that anticipatory programming for such scenarios is necessary for autonomous vehicles (unlike human drivers making emergency split-second decisions), and that such dilemmas are not to be taken literally as rare edge cases but rather as intuition pumps or thought experiments representing a wide range of pervasive harm–harm trade-offs. See, e.g., Jean-François Bonnefon, Azim Shariff & Iyad Rahwan, The Trolley, the Bull Bar, and Why Engineers Should Care About the Ethics of Autonomous Cars, 107 Proc. IEEE 502, 503 (2019) (arguing for the relevance of the moral reasoning generated by considering trolley problem dilemmas, despite their seeming unreality, for making practical decisions both about harm–harm trade-offs as well as "statistical trolley dilemma[s]," or risk–risk trade-offs); Bryan Casey, Amoral Machines, or: How Roboticists Can Learn to Stop Worrying and Love the Law, 111 Nw. U. L. Rev. Online 231, 239 (2017), http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1248&context=nulr_online [https://perma.cc/669R-3RL6] (arguing that "while society as a whole may remain agnostic to the trolley problem, engineers at the cutting edge of robotics are afforded no such luxury" because "trolley-like problems are not mere philosophical curiosities" but instead "are real-world contingencies that require prospective programming"); Samantha Godwin, Ethics and Public Health of Driverless Vehicle Collision Programming, 86 Tenn. L. Rev. 135, 143 (2018) (noting that "[s]omeone has to decide in advance what a driverless vehicle will do in situations where the vehicle cannot avoid a crash altogether but can select what object or person it will collide with"); Patrick Lin, Robot Cars and Fake Ethical Dilemmas, Forbes (Apr. 3, 2017), https://www.forbes.com/sites/patricklin/2017/04/03/robot-cars-and-fake-ethical-dilemmas/ [https://perma.cc/YP7Z-YX2L] (defending the relevance of self-driving-car dilemmas on various grounds).

[18] The surprise of the falling tree also limits a reflex people may have to say that Charlie should have anticipated the dangers ahead or should have been driving more slowly. That way of blaming Charlie would tend to bypass the presented dilemma.

[19] Or at least, to make any possible blame a bit more remote—say, if one wanted to blame the bicyclists for being there somehow, or to blame the car owner for putting yet another driverless car on the roads.

[20] The passenger is identified as the car's owner, however, both because that tends to be the relevant concern in the discourse and because it might be too easy for people to approve of sacrificing an unrelated passenger in lieu of two also-unrelated bicyclists. I also set the trade-off at two-to-one, to make it a bit of a harder case than the usual five-to-one trolley problem.

[21] It seemed a bit more plausible to say that the AI system could predict serious injuries than that it could predict certain death. Predicting injury rather than death also seemed a bit less likely to prompt the reader to infer that the car was going too fast to be safe (than if death were so easily predictable).

tree, is designed to blur several binaries that are usually central to trolley problems—distinctions between act versus omission, killing versus letting die, and intended effects versus side effects.[22] Downplaying these distinctions allows more of a focus on the desired contest between the principles of "protecting the passenger" and "minimizing casualties."

Third, I have left it to the subject's imagination how sophisticated Charlie is, aside from saying that it can drive a car on its own. In particular, the scenario does not say anything about Charlie's moral programming, nor about how Charlie "thinks" (beyond its ability to predict injuries). This is meant to approximate the degree of mystery or opacity that will naturally attend our interactions with such an advanced AI system; when driverless cars have become commonplace, how many people will really understand how the AI pilot learned to drive?[23] But I did name it Charlie, just as Alexa and Siri have names;[24] and I also describe Charlie as "ha[ving] to decide what to do." Both of these latter touches may enhance a tendency toward anthropomorphism in how some subjects view Charlie, though neither seems unusual even in the context of today's technology.[25]

[22] To elaborate: The car has already "acted" by swerving away from the tree in order to save the passenger; this muddies the intuition that the car would then merely be passive in continuing straight ahead to hit the bikers (letting them die, so to speak, by omission) rather than swerving again to sacrifice the passenger. Put differently, the initial swerving is intended to undermine the sense that there is one outcome that is obviously the natural course of things. (Is the natural course of things that the passenger remains safe after the car has initially swerved from the tree, or would it have been the natural course of things for the passenger to be injured by the crash with the tree?) The inclusion of an initial swerve may also help balance the two choices in terms of which injuries can be seen as an unintended side effect of saving someone else (as relevant to the doctrine of double effect): After the first swerve, it is possible to view the bicyclists' injuries (from the car) as a side effect of saving the passenger (from the tree), if Charlie drives the car forward; yet it is also possible to view the passenger's injuries (in the ditch) as a side effect of saving the bicyclists (from the car), if Charlie swerves a second time. For further discussion of these classic distinctions and how to blur them, see Huang, supra note 4, at 668–75 (describing philosopher Frances Kamm's ingenious example of a "bystanding driver" as well as the doctrine of double effect).

[23] It would be interesting in future work, however, to test how our moral judgments might be affected by various forms of "explainability"—for example, if Charlie were able to present a reason or justification for its decision. For explorations of explainability in this Symposium, see, for example, Ashley Deeks, The Judicial Demand for Explainable Artificial Intelligence, 119 Colum. L. Rev. 1829, 1830 (2019) (arguing that judges should play a leading role in developing rules for explainable AI); Katherine J. Strandburg, Rulemaking and Inscrutable Automated Decision Tools, 119 Colum. L. Rev. 1851, 1871–79 (2019) (assessing how the delegation and distribution of decisionmaking powers can complicate the role of explainability).

[24] As of the time of this study, the name "Charlie" did not seem to be associated with any well-known AI or machine-learning endeavor in the public eye.

[25] For example, it does not seem strange today to hear someone say, in a car, "Waze just changed its mind, because it sees an accident up ahead."

B. *Survey Population*

The survey subjects are adults living across the United States; they are volunteers recruited by the survey firm SurveyMonkey,[26] which approximated census-based age and gender distributions.[27] The sample analyzed below excludes anyone who did not complete the survey or who said that they could not take it seriously;[28] anyone who had taken another survey recently about a similar driverless car dilemma; anyone who failed a comprehension question about the scenario; and anyone who had attended law school or taken courses on artificial intelligence. The resulting sample includes 952 subjects, 59% of whom are women.

C. *Predictions*

What should we expect to see, in comparing across the three conditions, if telling people about the law influences their moral intuitions about Charlie's decision? It may be useful to consider two possible modes of such influence. First, knowing what the law says may exert a directional pull toward the moral judgments that align with the law's command and away from those that conflict with it. Second, if the law takes no stance, or if it expressly permits any action, then it may exert a pull toward moral neutrality and away from the moral poles.[29]

Various psychological mechanisms may interact to generate a directional influence from law's command, in ways that no doubt vary from person to person:[30] Some may see the law as a source of moral guidance,

[26] The subjects are volunteers in the sense that they are not paid for the time spent, nor paid a piece rate for each survey completed; rather, they are rewarded by SurveyMonkey in the following ways: A charitable donation is made for each survey completed, or they are entered in a sweepstakes. There is therefore a possibility that this survey population is somewhat more charitably inclined, or more interested in the offered sweepstakes program, than other people.

[27] Because the surveys were conducted online, however, the sample may lean toward those who have smart phones or home computers, and toward those who are comfortable with an online interface.

[28] A question expressly asked the subjects to confirm that they did take the survey seriously (or to say no), recognizing that everything in the scenarios was imaginary and set in the future.

[29] These two modes of influence were articulated in my prior work involving a human decisionmaker in classic trolley dilemmas. Huang, supra note 4, at 688–90 (describing the possibility of a "directional influence" and of a "pull of neutrality"). The discussion above, however, ignores two further possibilities. First, some people may hold a contrarian or oppositional attitude toward the law; and for any number of reasons, the law's instruction might generate reactance or backlash, boosting the opposed moral intuition or suppressing the aligned intuition. Second, there may be a crowding-out effect whereby the law's presence dampens the need for moral judgment. To the extent these seem plausible, one might interpret the observations in this study as net of such effects, but this study is unable to isolate them.

[30] This study is not designed to sort among these mechanisms or their interactions. Trying to do so would be a worthy, if perhaps daunting, aim for future extensions. For recent work critically surveying and contributing to various literatures about the

as a supplier of moral reasons, or as social proof of moral norms. Others may focus on a moral duty to obey the law, or they may view the costs of liability as morally relevant. For still others, the law might define social roles in a way that deserves moral respect, or it might more subtly affect moral psychology by implicitly establishing what is normal and what is a deviation.[31]

If such sources of the law's directional influence are at work, we might expect the {Protect passengers} condition to tip some subjects toward saying that Charlie's choice to sacrifice the passenger is "morally prohibited" and to tip some away from saying that it is "morally required."[32] Likewise, we might expect the {Minimize casualties} condition to tip some subjects toward saying that the decision is "morally required" and to tip some away from saying that it is "morally prohibited." Accordingly, the predictions for a comparison of the {Protect passengers} and {Minimize casualties} conditions are straightforward: The former should show more subjects saying that the sacrifice is "morally prohibited," and the latter should show more saying that it is "morally required."[33]

But what about comparisons with the {No laws} condition? This condition can be seen as a sort of neutral comparator, but with a cautious eye, as it is not an entirely natural baseline:[34] If subjects expect the law to

psychological internalization of law, see generally Yuval Feldman, The Law of Good People (2018); Richard H. McAdams, The Expressive Powers of Law (2015); Frederick Schauer, The Force of Law (2015); Bilz & Nadler, supra note 6.

[31] Defining what is normal, in this sense, may influence moral judgments through the act–omission distinction by setting an expectation about the natural course of things. As Professor Ronald Dworkin observed about the classic trolley problem:

> It is unclear what it means to let nature take its course. If it is natural to try to rescue five people at the cost of one, then throwing the switch is letting nature take its course. But perhaps "nature" means nonintelligent nature, so that a potential rescuer lets nature take its course by pretending that he is not there. But why should he?

Ronald Dworkin, Justice for Hedgehogs 298–99 (2011); cf. Adam Bear & Joshua Knobe, Normality: Part Descriptive, Part Prescriptive, 167 Cognition 25, 26–32 (2017) (providing experimental evidence for the argument "that people's normality judgments take into account both descriptive considerations (e.g., the statistical notion of the average) and more prescriptive considerations (e.g., what is morally ideal)").

[32] The expected effect of the {Protect passengers} condition on the share saying "morally permissible" could go either way, depending on whether more switch away from that category into "morally prohibited" or more switch in from "morally required." And the same reasoning applies, with switches in the other direction, for the {Minimize casualties} condition.

[33] As explained, the expected effect of either condition on the share saying "morally permissible" is ambiguous.

[34] Nonetheless, it still seems more informative for this thought experiment than other imaginable candidates for a neutral comparator. For example, an alternative might be to ask subjects to read the scenario without any mention of the law at all; yet that condition would be tricky to interpret as a baseline because its meaning would depend on what subjects are implicitly assuming about the background law (of the future)—a complication that is highlighted by this study's findings of law's influence on moral intuitions.

say *something* about safety priorities, at a time when driverless cars are common, then the {No laws} condition may suggest an implicit societal choice to deem any policy—prioritizing passengers, or minimizing casualties, or anything else—to be acceptable.[35] If so, the {No laws} condition may exert a pull of neutrality, drawing subjects away from the moral poles and toward the "morally permissible" answer. Comparisons of the {No laws} condition with the others, then, would be showing the net effects of law's directional influences and such a pull of neutrality.

II. Law's Influence

Telling people different things about the law leads to different moral judgments about Charlie's decision to sacrifice its sole passenger to save the two bicyclists. As seen in Table 1 and Figure 1, a comparison of the {Protect passengers} and the {Minimize casualties} conditions confirms the presence of law's directional pull.[36] The proportion of subjects who say that Charlie's decision is "morally prohibited" rises from 9% in the {Minimize casualties} condition to 16% in the {Protect passengers} condition.[37] Meanwhile, the proportion who say that the decision is "morally required" rises from 31% in the {Protect passengers} condition to 45% in the {Minimize casualties} condition.[38] Put differently, in the {Protect passengers} condition, there are about twice as many subjects saying the decision is "morally required" as saying "morally prohibited"; whereas in the {Minimize casualties} condition, that ratio increases to five times as many.

One might also compare each of these law conditions with the {No laws} condition—keeping in mind both that it is not an especially natural baseline and that the possibility of a pull of neutrality may complicate interpretation. Relative to {No laws}, the {Protect passengers} condition raises the share of subjects saying the decision is "morally prohibited" from 8% to 16%;[39] meanwhile, the share saying "morally required" does not seem to change by much.[40] Relative to the {No laws} condition, the

[35] This expectation may be lessened, however, by the fact that the survey subjects know they are taking a survey with an invented story—one that seems focused on the morality of the choice—and thus some may take the {No laws} statement as a cue to try to put any worry about the law out of their minds. This suggestion is purely speculative, of course.

[36] In a comparison of the {Protect passengers} and {Minimize casualties} conditions, there is no extra complication from the possibility of a pull of neutrality, as there will be in comparisons with the {No laws} condition.

[37] $\chi^2(1, N = 642) = 7.82, p = 0.005$. Pearson's chi-squared tests are reported throughout. Note that the standard error for any individual proportion $\hat{p}$ is given by the usual formula, the square root of $\hat{p}(1-\hat{p})$ / N.

[38] $\chi^2(1, N = 642) = 14.25, p < 0.001$.

[39] $\chi^2(1, N = 621) = 10.04, p = 0.002$.

[40] $\chi^2(1, N = 621) = 1.30, p = 0.254$. The standard reminder applies, throughout these analyses, that the lack of a statistically significant effect is not the same as evidence of zero effect.

{Minimize casualties} condition increases the share saying "morally required" from 35% to 45%;[41] meanwhile, the share saying "morally prohibited" seems about the same.[42] One possible explanation for these contrasts (between measurable change on one margin and a lack thereof on the other) is that some subjects are unmoved in their moral intuitions by hearing about a contrary law;[43] but another possibility is that a pull of neutrality in the {No laws} condition is masking the influence of the law conditions.[44]

TABLE 1: SACRIFICING THE PASSENGER

	Morally prohibited	Morally permissible	Morally required	*N*
Protect passengers	16.4%	53.1%	30.6%	311
Minimize casualties	9.1%	45.9%	45.0%	331
No laws	8.1%	57.1%	34.8%	310

[41] $\chi^2(1, N = 641) = 6.90, p = 0.009$.

[42] $\chi^2(1, N = 641) = 0.20, p = 0.652$.

[43] That is, it may be that those answering "morally required" in the {No laws} condition are not easily moved by hearing about the {Protect passengers} law, and those answering "morally prohibited" are not easily moved by hearing about the {Minimize casualties} law.

[44] To elaborate: Some subjects in the {Protect passengers} or {Minimize casualties} conditions, being no longer drawn toward the "morally permissible" answer by the pull of neutrality exerted by the {No laws} condition, may be inclined to choose one of the polar categories instead. Thus, in comparing the {No laws} and {Protect passengers} conditions, the observed shift toward "morally prohibited" might be due both to the law's directional influence and to relief from the pull of neutrality; whereas movement away from "morally required" might be offset by relief from the pull of neutrality. Likewise, in comparing the {No laws} and {Minimize casualties} conditions, the observed shift toward "morally required" might be due both to the law's directional influence and to relief from the pull of neutrality; whereas movement away from "morally prohibited" might be offset by relief from the pull of neutrality.

FIGURE 1A: SACRIFICING THE PASSENGER

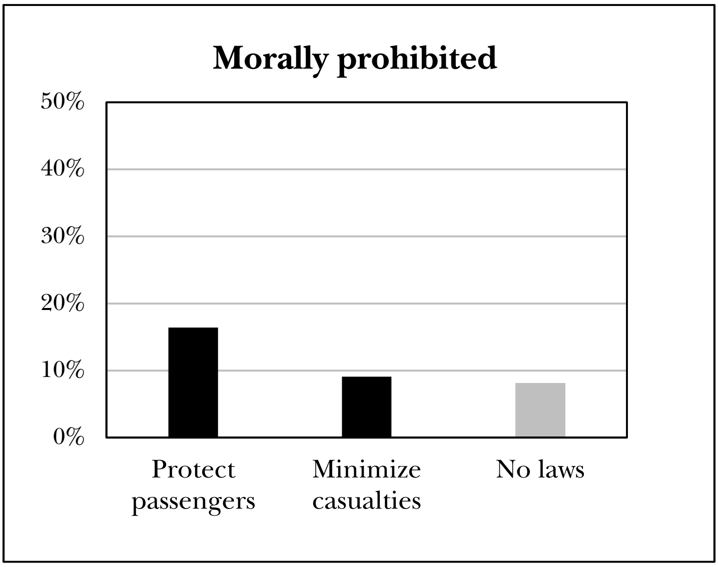
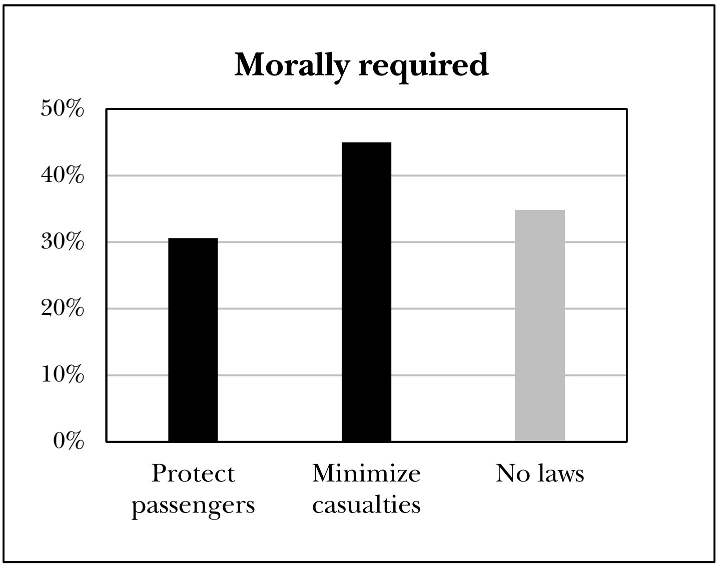


FIGURE 1B: SACRIFICING THE PASSENGER



A. *Whose Moral Responsibility?*

If someone says that Charlie's decision is "morally prohibited," whom do they blame? A follow-up question immediately after the main moral judgment question asks subjects to identify who "they think should bear some moral responsibility for that choice." Subjects may choose as many of the four given options as they see fit, and they may also choose "none of the above"; an open-ended text box invites them to suggest anyone else. Across the three conditions, 11% named Charlie as sharing in the moral responsibility; 66% named the company that created Charlie; 32% named the owner of the driverless car; 43% named the lawmakers; and 6% said none of the above.[45] (The shares do not add to one because the options are not mutually exclusive.)

These numbers should be taken with a grain of salt because the four answer options were preselected for the subjects. Still, looking at the *relative* attributions, two things seem to stand out:[46] First, lawmakers absorb a decent share of the assignment of moral responsibility—somewhat less than the company, but more than the car's owner. This finding speaks to a provocative question raised in this Symposium by Professor Tim Wu, who wonders whether in the near future the various defaults set by pervasive AI decisionmaking, through our acquiescence to them, will supersede the role of human-made law.[47] One might see this finding as a sign that the future public will not so readily let lawmakers off the hook, even when everyone knows that the AI systems will really be making the decisions. Or, from the perspective of the present, this finding may be seen as an expression of a generalized desire for lawmakers to deal with these hard questions before an AI does someday.[48]

[45] To be clear, these are shares of the subjects who say that Charlie's decision is "morally prohibited"—that is, those who morally object to the sacrifice of the passenger—across all three law conditions. This is not to ignore the distinction between moral blame and moral responsibility; for example, it would be coherent to ask subjects to assign moral responsibility for a blameless or a laudable decision (or even for a decision that has yet to be made).

[46] By relative attribution, I mean comparing the shares of people who assigned some moral responsibility to each of the four preselected actors. (The ordering of the four options is scrambled randomly for each subject.) Note, however, that these data do not tell us about the relative intensity of attribution. For example, a subject who names both Charlie and the car's owner might feel that the former bears less responsibility and the latter bears more; this difference would not appear in the data.

[47] As Professor Wu puts it: "Many of the developments that go under the banner of artificial intelligence that matter to the legal system are not so much new means of breaking the law but of bypassing it as a means of enforcing rules and resolving disputes." Tim Wu, Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems, 119 Colum. L. Rev. 2001, 2001 (2019) (citation omitted). The challenge to the legal system, then, would become the "existential challenge of supersession [by software]." Id.

[48] In concept it would be informative to consider each law condition separately, although in reality many more grains of salt must be taken, given the tiny sizes of the resulting subsamples—and again, we are only considering those subjects who answered "morally prohibited," the composition of whom differs across conditions. If one is willing

Second, these data suggest that people will not shy away from assigning moral responsibility to the creators of the AI—and for some people, even to the AI itself—for the choices made in such a dilemma.[49] Notably, among the options given, the company that created Charlie drew an attribution of moral responsibility from the most people. And even Charlie itself was named as morally responsible in part, by a small share of subjects. These relative attributions reflect how we currently envision the sophistication of such future AI pilots and how we imagine they might make such decisions.[50] As of now, it seems that relatively few subjects consider the AI itself to qualify as a moral agent. But in coming years, as we grow increasingly familiar with highly advanced and lifelike AI interfaces, one can only guess how our anthropomorphic imaginations might become more willing to assign moral agency to something—or will we say *someone*—like Charlie.

B. *Limitations*

Three sets of limitations are worth emphasizing. First, the standard disclaimer applies: What people say after reading a vignette might not reflect how they would react if similar events occurred in real life; and all the more so, for a story set in the future. In particular, this study's observations might be over- or understated, depending on various factors. For example, they might be overstated relative to a study in a natural setting (of future survey subjects) where people might not be paying much attention to the law. On the other hand, they might be understated relative to a future survey conducted after subjects have already had years of

to proceed with enough healthy skepticism: In the {Minimize casualties} condition, 50% of subjects assigned some moral responsibility to the lawmakers (N = 30); this may be unsurprising, as what the law says is also what Charlie actually does. In the {No laws} condition, 36% named the lawmakers (N = 25), which one might interpret as frustration for a failure of lawmaking. And yet there also seems to be ample attribution in the {Protect passengers} condition, with 43% naming the lawmakers as sharing in moral responsibility (N = 51), even though these subjects should be approving of the law, and even though Charlie's decision actually *violates* the law. To speculate, it thus seems possible that some sort of even more generalized attribution of moral responsibility to lawmakers is at work—perhaps for enabling such an AI-dominated state of the world in the first place.

[49] In their Symposium contributions, Professors Mala Chatterjee and Jeanne Fromer, as well as Professor Frank Pasquale, pose hard questions about the gaps in moral and legal accountability that might appear when we delegate decisionmaking to an AI. See Mala Chatterjee & Jeanne C. Fromer, Minds, Machines, and the Law: The Case of Volition in Copyright Law, 119 Colum. L. Rev. 1887, 1887 (2019) (discussing whether machines can be deemed to have mental states of the sort that would allow attributions of moral agency or of legal liability to them); Frank Pasquale, Data-Informed Duties in AI Development, 119 Colum. L. Rev. 1917, 1917 (2019) (discussing the potential for gaps in liability and responsibility when decisions are delegated to AI systems). This study's findings on the assignment of moral responsibility may be useful in calibrating such concerns, though of course it remains an open question how the future public will make such attributions.

[50] As noted above, the scenario intentionally says very little about how Charlie "thinks"—leaving this largely to the subject's imagination. See supra note 23 and accompanying text.

life experience with both driverless cars and the laws governing them. (That is to say, it may be expecting a lot of the present-day survey subjects to internalize an imaginary law of the future.) Moreover, the way the law is described in this study, without any mention of enforcement or liability, might also result in some understatement relative to a study with a more fully described legal regime—or relative to a future study done when those facts have become common knowledge.[51]

Second, this study is not designed to sort among the possible mechanisms of the law's halo. Do some subjects draw moral guidance or reasoning from what the law says? Do others see the law as a sign of society's moral norms? For some subjects, does the law implicitly set what is normal and what is deviant? For others, does it define social roles in a morally relevant way? Do some subjects place moral weight on the consequences of liability? Do others simply consider it a moral obligation to obey the law? Such psychological pathways may vary from person to person, and others may well be possible. Moreover, it goes without saying that while this study is designed to focus on the role of the law, many other sources of influence on our moral intuitions are worthy of study—and they may interact or interfere with the mechanisms of the law's halo in intriguing ways.

Third, there are limits of scope and generalizability. This survey includes only one scenario and focuses on only one decision by the AI; it also only considers two possible laws. Further work involving other variations of such dilemmas, and other possible laws or assignments of liability, is needed before any broad generalizations should be ventured.[52] Along these lines, a few suggestions follow.

CONCLUSION

This Essay presents an initial set of evidence suggesting that our moral intuitions about driverless car dilemmas can be influenced by the presence of the law. As this is but a single study, the most useful way to conclude may be to suggest a few dimensions of possible variation for

[51] See Huang, supra note 4, at 691–95 (showing evidence suggesting larger differences between contrasting law conditions when subjects are told that the deciding actor will be held liable than when subjects are told that there will be no liability despite what the law says). This study left liability open-ended in order to allow measuring of the subjects' natural tendencies in attributing moral responsibility. See supra section II.A. It would be an interesting extension to see if expressly assigning legal liability to a given actor might affect subjects' assignments of moral responsibility.

[52] While in prior work I have used survey experiments to explore the law's halo in a classic trolley dilemma, to my knowledge this study is the only one to have done so in the context of a driverless car dilemma. A different but related inquiry—about how the legal status of potential victims affects people's normative choices about whom to sacrifice (for example, whether people are less willing to save a jaywalker, or a criminal, relative to other potential victims)—has generated interesting findings in the MIT Moral Machine project. See Awad et al., supra note 8, at 59–63.

future work: First, different laws can be tested. For example, what if the law allows the car owner to control the "morality setting," dialing up or down the degree of priority for the passengers over outsiders? Second, other trade-offs can be posed. For example, the dilemma might involve increasing or decreasing risks, rather than causing harm for certain.[53] Third, the humans involved can be described with more morally relevant details. For example, what if the owner had bought this car believing that Charlie would put a priority on protecting passengers over protecting others? Or what if the passenger is in the driver's seat, and there *is* a steering wheel—but he declines to take over from the AI? And fourth, the AI's capabilities can be more fully specified. For example, what if it is known that Charlie learned to drive from watching millions of hours of videos of humans driving? Or what if Charlie is constrained to obey the law at all times? Or what if Charlie is not only the driver of a car, but an all-purpose companion who could easily pass the Turing test in conversation—even when discussing timeless moral dilemmas?

[53] For example, the car may be choosing whether to drive a bit to the left within its highway lane, closer to a motorcyclist on that side, and a bit more distant from a tractor-trailer on the right side; this trades off risk to the motorcyclist with risk to the car's passengers. See Bonnefon et al., supra note 17, at 503 (describing such a scenario as one example of a "statistical trolley dilemma"); cf. Barbara H. Fried, What *Does* Matter? The Case for Killing the Trolley Problem (or Letting It Die), 62 Phil. Q. 505, 506 (2012) (arguing that trolley-like thought experiments are limited in real-world relevance because they do not cover cases of "uncertain risk of accidental harm to generally unidentified others").

THE JUDICIAL DEMAND FOR EXPLAINABLE ARTIFICIAL INTELLIGENCE

*Ashley Deeks**

A recurrent concern about machine learning algorithms is that they operate as "black boxes," making it difficult to identify how and why the algorithms reach particular decisions, recommendations, or predictions. Yet judges are confronting machine learning algorithms with increasing frequency, including in criminal, administrative, and civil cases. This Essay argues that judges should demand explanations for these algorithmic outcomes. One way to address the "black box" problem is to design systems that explain how the algorithms reach their conclusions or predictions. If and as judges demand these explanations, they will play a seminal role in shaping the nature and form of "explainable AI" (xAI). Using the tools of the common law, courts can develop what xAI should mean in different legal contexts. There are advantages to having courts to play this role: Judicial reasoning that builds from the bottom up, using case-by-case consideration of the facts to produce nuanced decisions, is a pragmatic way to develop rules for xAI. Further, courts are likely to stimulate the production of different forms of xAI that are responsive to distinct legal settings and audiences. More generally, we should favor the greater involvement of public actors in shaping xAI, which to date has largely been left in private hands.

INTRODUCTION

A recurrent concern about machine learning algorithms is that they operate as "black boxes." Because these algorithms repeatedly adjust the way that they weigh inputs to improve the accuracy of their predictions, it can be difficult to identify how and why the algorithms reach the outcomes they do. Yet humans—and the law—often desire or demand answers to the questions "Why?" and "How do you know?" One way to address the "black box" problem is to design systems that explain how the algorithms reach their conclusions or predictions. Sometimes called "explainable AI" (xAI), legal and computer science scholarship has identified various actors who could benefit from (or who should demand) xAI. These include criminal defendants who receive long sentences based on opaque predictive algorithms,[1] military commanders who are

[*] E. James Kelly Jr.—Class of 1965 Research Professor of Law, University of Virginia Law School. Thanks to Danielle Citron, John Duffy, Bert Huang, Leslie Kendrick, Rich Schragger, and Rebecca Wexler for helpful comments, and to Scott Harman-Heath for excellent research assistance.

[1] See, e.g., Megan T. Stevenson & Christopher Slobogin, Algorithmic Risk Assessments and the Double-Edged Sword of Youth, 36 Behav. Sci. & L. 638, 639 (2018).

considering whether to deploy autonomous weapons,[2] and doctors who worry about legal liability for using "black box" algorithms to make diagnoses.[3] At the same time, there is a robust—but largely theoretical—debate about which algorithmic decisions require an explanation and which forms these explanations should take.

Although these conversations are critically important, they ignore a key set of actors who will interact with machine learning algorithms with increasing frequency and whose lifeblood is real-world controversies: judges.[4] This Essay argues that judges will confront a variety of cases in which they should demand explanations for algorithmic decisions, recommendations, or predictions. If and as they demand these explanations, judges will play a seminal role in shaping the nature and form of xAI. Using the tools of the common law, courts can develop what xAI should mean in different legal contexts, including criminal, administrative, and civil cases. Further, there are advantages to having courts play this role: Judicial reasoning that builds from the bottom up, using case-by-case consideration of the facts to produce nuanced decisions, is a pragmatic way to develop rules for xAI.[5] In addition, courts are likely to stimulate (directly or indirectly) the production of different forms of xAI that are responsive to distinct legal settings and audiences. At a more theoretical level, we should favor the greater involvement of public actors in shaping xAI, which to date has largely been left in private hands.

Part I of this Essay introduces the idea of xAI. It identifies the types of concerns that machine learning raises and that xAI may assuage. It then considers some forms of xAI that currently exist and discusses the advantages to each form. Finally, it identifies some of the basic xAI-related choices judges will need to make when they need or wish to understand how a given algorithm operates.

[2] See Matt Turek, Explainable Artificial Intelligence (XAI), DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence [https://perma.cc/ZNL9-86CF] (last visited Aug. 13, 2019).

[3] W. Nicholson Price II, Medical Malpractice and Black-Box Medicine, *in* Big Data, Health Law, and Bioethics 295, 295–96 (I. Glenn Cohen, Holly Fernandez Lynch, Effy Vayena & Urs Gasser eds., 2018).

[4] See, e.g., 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, 67 (2017) [hereinafter Edwards & Veale, Slave to the Algorithm] (questioning whether xAI will be useful because "[i]ndividual data subjects are not empowered to make use of the kind of algorithmic explanations they are likely to be offered" but ignoring the possible role for courts as users of xAI).

[5] Cf. Andrew Tutt, An FDA for Algorithms, 69 Admin. L. Rev. 83, 109 (2017) (proposing a federal statutory standard for explainability and arguing that "[i]f explainability can be built into algorithmic design, the presence of a federal standard could nudge companies developing machine-learning algorithms into incorporating explainability from the outset"). I share Andrew Tutt's view that it is possible to provide incentives for designers to incorporate xAI into their products, but I believe that there are advantages to developing these rules using common law processes.

Against that background, the Essay then turns to two concrete areas of law in which judges are likely to play a critical role in fleshing out whether xAI is required and, if so, what forms it should take. Part II considers the use of machine learning in agency rulemaking and adjudication and argues that judges should insist on some level of xAI in evaluating the reasons an agency gives when it produces a rule or decision using algorithmic processes.[6] Further, if agencies employ advanced algorithms to help them sort through high volumes of comments on proposed rules, judges should seek explanations about those algorithms' parameters and training.[7] In both cases, if judges demand xAI as part of the agency's reason-giving process, agency heads themselves will presumably insist that their agencies regularly employ xAI in anticipation of litigation.

Part III explores the use of predictive algorithms in criminal sentencing. These algorithms predict the likelihood that a defendant will commit additional crimes in the future. Here, the judge herself is the key consumer of the algorithm's recommendations, and has a variety of incentives—including the need to give reasons for a sentence, concerns about reversal on appeal, a desire to ensure due process, and an interest in demonstrating institutional integrity—to demand explanations for how the sentencing algorithm functions.

As courts employ and develop existing case law in the face of predictive algorithms that arise in an array of litigation, they will create the "common law of xAI," law sensitive to the requirements of different audiences (judges, juries, plaintiffs, or defendants) and different uses for the explanations given (criminal, civil, or administrative law settings).[8] A nuanced common law of xAI will also provide important incentives and feedback to algorithm developers as they seek to translate what are currently theoretical debates into concrete xAI tools.[9] Courts should focus on the power of xAI to identify algorithmic error and bias and the need

[6] For the argument that judicial review of agency rulemaking employs common law methodologies, see Jack M. Beermann, Common Law and Statute Law in Administrative Law, 63 Admin. L. Rev. 1, 3 (2011).

[7] See Melissa Mortazavi, Rulemaking Ex Machina, 117 Colum. L. Rev. Online 202, 207–08 (2017), https://columbialawreview.org/wp-content/uploads/2017/09/Mortavazi-v5.0.pdf [https://perma.cc/SF8R-EG9C] (examining the possibility that agencies may deploy automated notice-and-comment review).

[8] See Finale Doshi-Velez & Mason Kortz, Berkman Klein Ctr. Working Grp. on Explanation & the Law, Accountability of AI Under the Law: The Role of Explanation 12 (2017), https://arxiv.org/pdf/1711.01134.pdf [https://perma.cc/LQB3-HG7L] ("As we have little data to determine the actual costs of requiring AI systems to generate explanations, the role of explanation in ensuring accountability must also be re-evaluated from time to time, to adapt with the ever-changing technology landscape.").

[9] At least one scholarly piece has concluded that "there is some danger of research and legislative efforts being devoted to creating rights to a form of transparency that may not be feasible, and may not match user needs." Edwards & Veale, Slave to the Algorithm, supra note 4, at 22. A common law approach to xAI can help ensure that the solutions are both feasible and match user needs in specific cases.

for xAI to be comprehensible to the relevant audience. Further, they should be attuned to dynamic developments in xAI decisions across categories of cases when looking for relevant precedent and guidance.

I. THE WHAT AND WHY OF EXPLAINABLE AI

Artificial intelligence is a notoriously capacious and slippery term. Generally, it refers to "a set of techniques aimed at approximating some aspect of human or animal cognition using machines."[10] More concretely, scientists and scholars often use the term to encompass technologies that include machine learning, speech recognition, natural language processing, and image recognition.[11] Machine learning systems and algorithms, the driving force behind many AI developments, are valuable because of their ability to learn for themselves "how to detect useful patterns in massive data sets and put together information in ways that yield remarkably accurate predictions or estimations."[12] Many machine learning systems are trained on large amounts of data and adjust their own parameters to improve the reliability of their predictions over time.[13] Machine learning tools hold out the possibility of making more accurate decisions, faster, based on far larger quantities of data than humans can process and manipulate.[14] Importantly, though, because a machine learning system learns on its own and adjusts its parameters in ways its programmers do not specifically dictate, it often remains unclear precisely how the system reaches its predictions or recommendations.[15] This is particularly true for "deep learning" systems that use "neural networks," which are intended to replicate neural processes in the human brain.[16] Deep learning systems use nodes, arranged in multiple layers, which transfer information to each other and learn on their own how to weigh

[10] Ryan Calo, Artificial Intelligence Policy: A Primer and Roadmap, 51 U.C. Davis L. Rev. 399, 404 (2017).

[11] Artificial Intelligence, Lexico, https://www.lexico.com/en/definition/artificial_intelligence [https://perma.cc/MNB4-ZENF] (last visited Oct. 15, 2019) (defining "artificial intelligence" as "[t]he theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages").

[12] Cary Coglianese & David Lehr, Transparency and Algorithmic Governance, 71 Admin. L. Rev. 1, 6 (2019) [hereinafter Coglianese & Lehr, Governance]; see also id. at 14–16 (describing how machine learning differs from traditional statistical techniques).

[13] See Ethem Alpaydin, Machine Learning: The New AI 24–25 (2016).

[14] See Coglianese & Lehr, Governance, supra note 12, at 16.

[15] See Alpaydin, supra note 13, at 155; Will Knight, The Dark Secret at the Heart of AI, MIT Tech. Rev., May/June 2017, at 55, 56–57.

[16] See James Farrant & Christopher M. Ford, Autonomous Weapons and Weapon Reviews: The UK Second International Weapon Review Forum, 93 Int'l L. Stud. 389, 400 (2017); see also David Lehr & Paul Ohm, Playing with the Data: What Legal Scholars Should Learn About Machine Learning, 51 U.C. Davis L. Rev. 653, 693 & n.135 (2017).

connections between nodes to correctly interpret objects in, say, a video image.[17]

Notwithstanding its potential benefits, the use of machine learning has prompted a number of concerns, especially when the systems make predictions that affect people's liberty, safety, or privacy. One strand of criticism focuses on the ways in which these algorithms can replicate and exacerbate societal biases in light of the data on which scientists train them. Another line of critiques questions the accuracy of various machine learning predictions, with objectors claiming that tools such as criminal justice algorithms predict recidivism less accurately than humans.[18]

A third concern, and the one most salient to this Essay, centers on the lack of information about how the algorithm arrives at its results—the "black box" problem.[19] The inability to parse the reasons behind the algorithm's recommendations can harm those affected by the recommendations. Opaque algorithms can undercut people's sense of fairness and trust—particularly when used by the government—and in the criminal justice setting can undercut a defendant's right to present a defense. This Essay focuses on algorithms' lack of transparency and interpretability for two related reasons. First, shedding light on how an algorithm produces its recommendations can help address the other two critiques, by allowing observers to identify biases and errors in the algorithm.[20] Second, computer scientists have begun to make promising inroads into the problem by developing what is often referred to as "explainable AI."[21]

[17] See Farrant & Ford, supra note 16, at 400–01.

[18] See Julia Dressel & Hany Farid, The Accuracy, Fairness, and Limits of Predicting Recidivism, Sci. Advances, Jan. 2018, at 1, 3, https://advances.sciencemag.org/content/4/1/eaao5580/tab-pdf (on file with the *Columbia Law Review*).

[19] See, e.g., Frank Pasquale, The Black Box Society: The Secret Algorithms that Control Money and Information 3–4 (2015); Danielle Keats Citron, Technological Due Process, 85 Wash. U. L. Rev. 1249, 1254 (2008) (expressing concern about the "opacity of automated systems" used to inform administrative rulemaking).

[20] Finale Doshi-Velez & Been Kim, Towards a Rigorous Science of Interpretable Machine Learning 1, 3 (2017), https://arxiv.org/pdf/1702.08608.pdf [https://perma.cc/ALR4-DM7J] ("[I]f the system can *explain* its reasoning, we then can verify whether that reasoning is sound with respect to . . . other desiderata—such as fairness, privacy, reliability, robustness, causality, usability and trust . . . .").

[21] For a recent survey of developments in xAI, see Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter & Lalana Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, arXiv (May 31, 2018), https://arxiv.org/pdf/1806.00069.pdf [https://perma.cc/3SG4-G5GA] (last updated Feb. 3, 2019). One reason for recent progress in this area is the entry into force of the European Union's General Data Protection Regulation, which contains provisions that arguably give individuals affected by purely algorithmic decisions a "right to an explanation." See Regulation (EU) 2016/679 of the European Parliament and of 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, arts. 13–15, 2016 O.J. (L 119) 41–43 (providing rights to "meaningful information about the logic involved" in certain automated decisions). The existence of these provisions, coupled with a lack of detail about what form those explanations must take, has triggered extensive discussions in

xAI encompasses a range of efforts to explain—or help humans interpret—how a particular machine learning model reached its conclusion. The concept of an explanation here "has come to refer to providing insight into the internal state of an algorithm, or to human-understandable approximations of the algorithm."[22] xAI provides a variety of benefits: It can foster trust between humans and the system,[23] identify cases in which the system appears to be biased or unfair, and bolster our own knowledge of how the world works.[24] As discussed below, in legal settings xAI can benefit judges who wish to rely on the algorithms for decisional support, litigants who seek to persuade judges that their use of algorithms is defensible, and defendants who wish to challenge predictions about their dangerousness.[25] xAI is not without costs, however. Most significantly, making an algorithm explainable may result in a decrease in its accuracy.[26] xAI may also stifle innovation, force developers to reveal trade secrets, and impose high monetary costs because xAI can be expensive to build.[27]

Fortunately, a variety of xAI currently exists, and computer scientists continue to develop new forms of it.[28] Some machine learning models are built to be intrinsically explainable, yet these models are often less

the legal and machine learning communities about how and in what form to explain the results of highly complex algorithms to experts and nonexperts.

[22] Sandra Wachter, Brent Mittelstadt & Chris Russell, Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, 31 Harv. J.L. & Tech. 841, 850 (2018); see also Doshi-Velez & Kim, supra note 20, at 2 (defining interpretability as the "ability to explain or to present in understandable terms to a human").

[23] See Knight, supra note 15, at 61 (describing "explainability as the core of the evolving relationship between humans and intelligent machines"); Turek, supra note 2 ("Explainable AI—especially explainable machine learning—will be essential if future warfighters are to understand, appropriately trust, and effectively manage an emerging generation of artificially intelligent machine partners.").

[24] See Doshi-Velez & Kim, supra note 20, at 3.

[25] See, e.g., Robin A. Smith, Opening the Lid on Criminal Sentencing Software, Duke Today (July 19, 2017), https://today.duke.edu/2017/07/opening-lid-criminal-sentencing-software [https://perma.cc/F63A-VWLQ] ("Using . . . machine learning, Rudin and colleagues are training computers to build statistical models to predict future criminal behavior . . . that are just as accurate as black-box models, but more transparent and easier to interpret.").

[26] See Doshi-Velez & Kortz, supra note 8, at 2 ("[E]xplanation would come at the price of system accuracy or other performance objective[s].").

[27] See id. at 2, 12 ("Requiring every AI system to explain every decision could result in less efficient systems, forced design choices, and a bias towards explainable but suboptimal outcomes.").

[28] This Essay's discussion of categories of xAI is necessarily simplified, because there are a wide range of approaches to categorizing xAI and the nomenclature is unsettled. For a survey of the literature on types of xAI and a detailed taxonomy thereof, see Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi & Fosca Giannotti, A Survey of Methods for Explaining Black Box Models 6–8 (2018), https://arxiv.org/pdf/1802.01933.pdf [https://perma.cc/P8PH-Z5V7].

complex as a result and tend to be less accurate in their predictions.[29] Another set of models is not intrinsically explainable. For these models, computer scientists have taken two basic approaches.[30] One type (which this Essay terms an "*exogenous* approach") does not attempt to actually explain the inner workings of (that is, the reasoning of) the machine learning algorithm. Instead, it attempts to provide relevant information to the algorithm's user or subject about how the model works using extrinsic, orthogonal methods.[31] A second type of approach actually attempts to explain or replicate the model's reasoning, and sometimes is referred to as a "*decompositional* approach."[32]

Exogenous xAI approaches can either be *model-centric* or *subject-centric*.[33] A model-centric approach, also referred to as global interpretability,[34] might involve, for instance, explaining the creator's intentions behind the modelling process, the family of model the system uses, the parameters the creators specified before training the system, qualitative descriptions of the input data the creator used to train the model, how the model performed on new data, and how the creators tested the data for undesirable properties.[35] In other words, this constitutes a thick description of the parts of the model that are knowable. A different type of model-centric approach might audit the outcomes of the machine learning system.[36] This approach would scour the system's decisions or recommendations for appearances of bias or error. Model-centric approaches attempt to explain the whole model, rather than its performance

[29] These include linear, parametric, and tree-based models. Dipanjan Sarkar, The Importance of Human Interpretable Machine Learning, Towards Data Sci. (May 24, 2018), https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476 [https://perma.cc/4XD8-F7CD]. For an argument that society should use only intrinsically interpretable models for high-stakes decisions, see generally Cynthia Rudin, Please Stop Explaining Black Box Models for High-Stakes Decisions (2018), https://arxiv.org/pdf/1811.10154.pdf [https://perma.cc/Q7SF-6DYN].

[30] See Guidotti et al., supra note 28, at 2 (characterizing one category of xAI as focused on describing how black boxes work and another on explaining decisions without understanding how the decision systems work); Edwards & Veale, Slave to the Algorithm, supra note 4, at 64–65 (describing two styles of algorithmic explanation: one that "opens" the black box and one that does not).

[31] See Edwards & Veale, Slave to the Algorithm, supra note 4, at 65 ("[P]edagogical systems . . . can get the information they need by simply querying it, like an oracle." (emphasis omitted)).

[32] Id. at 64.

[33] Id. at 22.

[34] See Sarkar, supra note 29.

[35] See Edwards & Veale, Slave to the Algorithm, supra note 4, at 55–56.

[36] Joshua A. Kroll, Solon Barocas, Edward W. Felton, Joel R. Reidenberg, David G. Robinson & Harlan Yu, Accountable Algorithms, 165 U. Pa. L. Rev. 633, 660–61 (2017) (explaining that auditing may test for discrimination in bargaining processes such as retail car negotiations).

in a particular case, and can help ensure that decisions are being made in a procedurally regular way.[37]

A subject-centric approach, also referred to as local interpretability,[38] in contrast, might provide the subject of a recommendation or decision with information about the characteristics of individuals who received similar decisions.[39] Another subject-centric approach involves the use of counterfactuals.[40] Here, people seeking to understand which factors may have most affected the algorithm's recommendation about them may, using that same algorithm, tweak the input factors to test how much a given factor mattered in the original recommendation.[41] For example, an algorithm that deems someone convicted of an offense to be at high risk of reoffending could be tested with counterfactuals to see whether the recommendation would have been different if the person were ten years older, or had one fewer arrest. The counterfactual approach could take different forms: It might present several "close possible worlds" or one "closest possible world," and it might alter one factor or several different factors.[42] One advantage of an exogenous approach is that it does "not require the data subject to understand any of the internal logic of a model in order to make use of it."[43] Subject-centric approaches can be particularly useful for individuals who are seeking to understand "if and how they might achieve a different outcome"; they empower an individual

[37] See Edwards & Veale, Slave to the Algorithm, supra note 4, at 55–56.

[38] See Sarkar, supra note 29 (defining local interpretability as trying to understand why the model made a particular decision in a single instance).

[39] See Edwards & Veale, Slave to the Algorithm, supra note 4, at 58.

[40] See Wachter et al., supra note 22, at 845 ("In the existing literature, 'explanation' typically refers to an attempt to convey the internal state or logic of an algorithm that leads to a decision. In contrast, counterfactuals describe a dependency on the external facts that led to that decision.").

[41] See id. at 854, 881–82 (discussing implementation options); see also Danielle Keats Citron & Frank Pasquale, The Scored Society: Due Process for Automated Predictions, 89 Wash. L. Rev. 1, 28–29 (2014) (proposing a system to allow consumers to enter "hypothetical alterations" to their credit histories to see how the alterations affect their score).

[42] See Wachter et al., supra note 22, at 848 ("Such considerations [relevant to which type of counterfactual you produce] may include the capabilities of the individual concerned, sensitivity, mutability of the variables involved in a decision, and ethical or legal requirements for disclosure."); id. at 851 (noting that one could offer "multiple diverse counterfactual explanations to the data subject"); see also Edwards & Veale, Slave to the Algorithm, supra note 4, at 63 (describing how counterfactual models can allow individuals to view and reflect upon the decisions about other users).

[43] Wachter et al., supra note 22, at 851; id. at 860 ("[C]ounterfactuals bypass the substantial challenge of explaining the internal workings of complex machine learning systems," providing information that "is both easily digestible and practically useful for understanding the reasons for a decision, challenging them, and altering future behaviour for a better result.").

to more effectively navigate and challenge the process in a particular case.[44]

An alternative to these exogenous approaches is a category of xAI that attempts to explain (or "decompose") the model's reasoning. The most obvious way to do so is to reveal the source code for the machine learning model, but that approach will often prove unsatisfactory (because of the way machine learning works and because most people will not be able to understand the code).[45] More nuanced alternatives exist, however. One approach is to create a second system alongside the original "black box" model, sometimes called a "surrogate model."[46] A surrogate model works by analyzing featured input and output pairs but does not have access to the internal weights of the model itself.[47] For instance, scholars constructed a decision tree that effectively mirrored the computations of a black box model that predicted patients' risk for diabetes. The decision tree allowed computer scientists to track which factors (such as cholesterol level, nicotine dependence, and edema) the black box model weighed in making its risk assessments.[48] In a legal setting, this approach might entail creating a decision tree that accurately reconstructs the decisions of a self-driving car's black box algorithms in a product liability case, for example. These systems closely approximate the predictions made by an underlying model, while being interpretable.[49]

There are a host of ways in which machine learning algorithms will find their way into court in coming years. As a result, the courts themselves will be important actors in the machine learning ecosystem that is working to decide when, how, and in what form to develop xAI for algorithms. In specific cases, courts will need to consider a range of questions: Who is the audience for the explanation, and how simple or complex should the explanation be? How long should it take the user to understand the explanation?[50] What structure or form should the xAI take: lines of code, visual presentations, manipulable programs?[51] What

[44] See Andrew D. Selbst & Solon Barocas, The Intuitive Appeal of Explainable Machines, 87 Fordham L. Rev. 1085, 1120 (2018).

[45] Kroll et al., supra note 36, at 638–39 (arguing that revealing source code is a misguided way of creating algorithmic accountability).

[46] W. Andrew Pruett & Robert L. Hester, The Creation of Surrogate Models for Fast Estimation of Complex Model Outcomes, PLOS One (June 3, 2016), https://doi.org/10.1371/journal.pone.0156574 [https://perma.cc/GZ33-78MC].

[47] Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, arXiv (Aug. 9, 2016), https://arxiv.org/pdf/1602.04938.pdf [https://perma.cc/8WN8-WQJF].

[48] Osbert Bastani, Carolyn Kim & Hamsa Bastani, Interpreting Blackbox Models via Model Extraction, arXiv (Jan. 24, 2019), https://arxiv.org/pdf/1705.08504.pdf [https://perma.cc/L2K4-ZPVU].

[49] See id.

[50] See Doshi-Velez & Kim, supra note 20, at 7–8.

[51] See Wachter et al., supra note 22, at 872 (noting that one could disclose the algorithm's source code, formula, weights, and full set of variables).

factors should the explanation focus on? When should xAI be model-centric and when should it be subject-centric? If there are trade secrets at issue, should the court review the algorithm in camera or request an independent peer review under a nondisclosure agreement?[52] More generally, what will constitute a "meaningful explanation"?[53] Judges are well positioned in this ecosystem to develop pragmatic approaches to xAI, even though they are not—indeed, *because* they are not—experts in machine learning technology.

To understand how these questions may arise concretely in practice, the next Parts identify and analyze two legal settings in which courts soon will need to make decisions about the types of xAI that are helpful—or that may even be legally required.

II. ALGORITHMS IN AGENCY RULEMAKING AND ADJUDICATION

Scholars have begun to consider the ways in which machine learning algorithms could advance the work of administrative agencies.[54] Cary Coglianese and David Lehr write, "[N]ational security and law enforcement agencies are starting to rely on machine learning . . . . [O]ther government agencies have also begun to explore uses of machine learning, revealing growing recognition of its promise across a variety of policy settings and at all levels of government."[55] Machine learning algorithms

[52] See Coglianese & Lehr, Governance, supra note 12, at 49 (suggesting these two methods of review as ways to balance the need for transparency in administrative decision-making with the need to protect trade secrets).

[53] In the national security context, judges frequently have to decide what types of classified explanations by the executive branch are sufficient. See Ashley S. Deeks, Secret Reason-Giving, 129 Yale L.J. (forthcoming 2019) (manuscript at 22–24) (on file with the *Columbia Law Review*) (discussing secret reason-giving in the context of foreign surveillance, asset freezes, state secrets, and the Freedom of Information Act).

[54] See generally, e.g., Citron, supra note 19 (expressing concern that automated decisionmaking will undermine procedural safeguards and displace expert reasoning); Cary Coglianese & David Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era, 105 Geo. L.J. 1147 (2017) [hereinafter Coglianese & Lehr, Regulating by Robot] (arguing that the use of machine learning in administrative actions does not violate the nondelegation doctrine, due process, equal protection, or the reasoned explanation requirements of the Administrative Procedure Act); Mariano-Florentino Cuéllar, Cyberdelegation and the Administrative State, *in* Administrative Law from the Inside Out: Essays on Themes in the Work of Jerry L. Mashaw 134 (Nicholas R. Parillo ed., 2017) (highlighting the potential tradeoff between the increased precision of artificial intelligence decisionmaking and the risk of displacing agency deliberation about social welfare); Benjamin Alarie, Anthony Niblett & Albert Yoon, Regulation by Machine (Dec. 1, 2016) (unpublished manuscript), https://ssrn.com/abstract =2878950 (on file with the *Columbia Law Review*) (envisioning that agencies may deploy algorithms to predict how courts will decide administrative law cases).

[55] Coglianese & Lehr, Regulating by Robot, supra note 54, at 1161; see also Coglianese & Lehr, Governance, supra note 12, at 3 ("Scholars and policy officials alike see increasing promise for the use of machine-learning algorithms by administrative agencies in a range of domestic policy areas.").

offer the potential to support agency rulemaking and also perhaps adjudications.[56] Virtually all of the scholars who have studied the issue anticipate that agencies' use of algorithms will only increase in coming years.[57]

Consider how agencies might deploy machine learning algorithms to facilitate rulemaking. Justice Mariano-Florentino Cuéllar writes, "Over time, neural networks and genetic algorithms will almost certainly inform judgments about the proper scope of a rule . . . ."[58] Coglianese and Lehr go further, envisioning truly autonomous rulemaking in areas such as SEC regulation of high-speed electronic trading or Treasury Department regulations that respond to real-time market changes suggestive of systemic risk.[59] They even envision multiagent systems, where machine learning algorithms would model different forecasts for different values to be traded off, and a separate machine learning system representing the agency would pick the model (and hence the rule) that maximizes the objective selected by humans.[60]

Another opportunity for the use of machine learning algorithms in the agency setting might be to parse and summarize voluminous public comments provided as part of notice and comment rulemaking.[61] Further, as noted above, agencies might turn to machine learning to help them conduct adjudications.[62] This could include using algorithms to predict pilot competence and grant pilot's licenses, forecast the effects of a proposed merger on competition, or decide disability claims.[63] None of these processes will exclude the human role entirely—at the very least, computer scientists must code agency "values" into the algorithms in the form of ones and zeros—but machine learning–driven rulemaking and adjudication may embody a host of decisional steps that are nontransparent and difficult to trace.

Courts are likely to confront all of these agency uses of algorithms. Under the Administrative Procedure Act (APA), courts generally may

[56] Coglianese & Lehr, Regulating by Robot, supra note 54, at 1167 (discussing possible applications of machine learning in administrative rulemaking and adjudications).

[57] See, e.g., Cuéllar, supra note 54, at 135 ("Reliance on computer programs to make administrative decisions — whether designed as conventional expert systems, more elaborate genetic or otherwise self-modifying algorithms, neural or 'deep learning' networks, or other machine learning mechanisms — will likely accelerate.").

[58] Id. at 144.

[59] Coglianese & Lehr, Regulating by Robot, supra note 54, at 1171–72.

[60] Id. at 1174; Coglianese & Lehr, Governance, supra note 12, at 9–10; see also Cuéllar, supra note 54, at 17.

[61] Mortazavi, supra note 7, at 207–08.

[62] See Coglianese & Lehr, Governance, supra note 12, at 9 (noting that "the statistical tools that will facilitate adjudicating by algorithm already exist and are already being employed in analogous endeavors"); Cuéllar, supra note 54, at 137 (envisioning "sleek black boxes" administering "bureaucratic justice").

[63] Coglianese & Lehr, Regulating by Robot, supra note 54, at 1170–71; Cuéllar, supra note 54, at 136–37.

review final agency actions.[64] For example, in the informal rulemaking context, courts may review agency factual determinations and discretionary decisions and set aside those actions that are arbitrary, capricious, or an abuse of discretion.[65] In that context, the Supreme Court requires an agency to "examine the relevant data and articulate a satisfactory explanation for its action including a 'rational connection between the facts found and the choice made.'"[66] More recently, the Court confirmed that the courts' role involves "examining the *reasons* for agency decisions—or, as the case may be, the absence of such reasons."[67] Agencies also are expected to address salient points raised in public comments.[68] That said, courts will give an agency particular deference when the agency is making predictions within its area of expertise that involve technical matters ("at the frontiers of science").[69]

Agency reason-giving thus plays an important role in defending the rules that agencies produce. Yet reason-giving can be complicated, if not confounded, by machine learning algorithms. An agency that has relied heavily on a machine learning algorithm prediction about the impact of a particular chemical on human health or about the population trajectory of a threatened species may need to share with the court the types of data it used, the type of machine learning model it used, the algorithm's error rate, and—possibly—the way the algorithm functioned to produce its prediction.[70] It is not yet clear precisely what courts will demand of agencies in this setting, or how agencies will respond.

Some scholars are relatively sanguine about the ease with which courts will adjust to the growing use of algorithms by agencies. For example, Coglianese and Lehr argue that current legal standards in administrative law do not demand anything close to transparency, that courts apply a deferential standard to agency rulemaking that relies on complex modelling, and that agencies will generally be able to meet that standard if they can show that the algorithm has performed as intended and achieves a justified objective.[71] Other scholars are more skeptical. Danielle Citron, for instance, worries that opaque algorithms impair meaningful judicial review because courts cannot see the rules that are

[64] 5 U.S.C. § 702 (2012); Block v. Cmty. Nutrition Inst., 467 U.S. 340, 345 (1984).

[65] 5 U.S.C. § 706(2)(A); see also id. § 553.

[66] Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 43 (1983) (quoting Burlington Truck Lines, Inc. v. United States, 371 U.S. 156, 168 (1962)).

[67] Judulang v. Holder, 565 U.S. 42, 53 (2011) (emphasis added).

[68] Perez v. Mortg. Bankers Ass'n, 135 S. Ct. 1199, 1203 (2015) ("An agency must consider and respond to significant comments received during the period for public comment.").

[69] Balt. Gas & Elec. Co. v. NRDC, 462 U.S. 87, 103 (1983).

[70] See Cuéllar, supra note 54, at 151–52 (noting that courts may want to understand how that process occurred and how users tested the system to ensure those values were fairly captured in the output).

[71] See Coglianese & Lehr, Governance, supra note 12, at 35–36, 39, 47–49.

actually applied in a given case.[72] One possibility is that a court could reduce its level of deference to an agency decision when the agency deploys a black-box algorithm purchased from the private sector, because the court concludes that the agency is making a prediction based on private sector expertise, not its own.

Whether optimistic or pessimistic about the way courts will address these challenges, many scholars take comfort in xAI's possibilities. Justice Cuéllar contemplates that machine learning algorithms may help agencies withstand judicial scrutiny, because he assumes that their use could "conceivably yield greater transparency by making it easier to follow what precise considerations were used in driving a particular outcome."[73] This is only true, of course, if some form of xAI accompanies the algorithm. Likewise, Coglianese and Lehr admit that xAI will make it easier to defend an extensive use of machine learning algorithms by agencies. They highlight the "widening panoply of techniques that data scientists are developing to make learning algorithms more explainable" and note that even when the government uses algorithms to make individual-level predictions, "government agencies will likely have strategies available to them to provide individual-level explanations."[74] In short, xAI is likely to serve as an important linchpin in agencies' transition from human-dominated decisionmaking to machine-dominated decisionmaking. Yet none of these scholars focus on the direct role that the courts will play in affecting xAI itself.

As courts work through administrative law cases involving machine learning algorithms, they will play a significant role in shaping the xAI ecosystem. The extent to which courts seek information about the inputs, outputs, and reliability of agency algorithms or express interest in testing counterfactuals will give concrete form to current xAI discussions, which are happening largely in the abstract. Courts' approaches to agency algorithms in rulemaking settings might prompt developers to pursue exogenous xAI approaches, using model-centric explanations to defend the overall workings and reliability of the algorithm. Courts' approaches to agency algorithms in adjudication, in contrast, might lead developers to pursue decompositional approaches, using subject-centric explanations to defend the specific adjudicatory choices made. The healthy and growing set of xAI tools means that there is a range of choices from which to draw—and, as of now, no statutory guidance about xAI.

The prospect of courts being able to select the proper xAI tool for a given situation is a good thing, for all of the reasons that we celebrate the

[72] Citron, supra note 19, at 1298.

[73] Cuéllar, supra note 54, at 142, 153. Cuéllar seems less sanguine about situations in which xAI is not available, noting with concern that decisions could "be made on a basis phenomenologically different from what could easily be understood or even explained by human participants." Id. at 157.

[74] Coglianese & Lehr, Governance, supra note 12, at 6, 55.

strengths of the common law.[75] Courts can move "cautiously and incrementally" as they sort out what types of xAI will be effective and realistically achievable in explaining different types of agency algorithms.[76] The courts will confront a set of concrete facts, and can, as a result, produce context-sensitive holdings that do not attempt to impose broad policies on xAI developments. Further, the courts here will build on existing case law that fleshes out the requirements of the APA, modestly adjusting that case law for situations in which the use of this new technology raises unanswered questions.[77]

xAI may also mitigate changes in the law that otherwise could result from the technological disruptions wrought by machine learning. For example, if courts become concerned about continuing to accord deference to agency decisionmakers who rely heavily on algorithms or worry about granting opaque algorithmic decisionmaking a "presumption of regularity,"[78] xAI may help assuage these concerns. Agencies may perceive the advantages of adopting xAI as a means to address judicial concerns ex ante and thus to minimize disadvantageous doctrinal changes.[79] Although common law xAI will, at least initially, offer less predictability than a federal xAI statute would, it can more easily take into account technological developments in xAI, and it can be more sensitive to what is both necessary and possible in a given setting.

III. CRIMINAL SENTENCING ALGORITHMS

In the administrative law setting, judges will sit as neutral reviewers of an agency's use of machine learning algorithms. In the criminal justice setting, judges themselves may be the ones using those algorithms.[80]

[75] For a general discussion of the advantages of developing rules through the common law rather than by statute, see generally Jeffrey J. Rachlinski, Bottom-Up Versus Top-Down Lawmaking, 73 U. Chi. L. Rev. 933 (2006).

[76] Neal Devins & David Klein, The Vanishing Common Law Judge?, 165 U. Pa. L. Rev. 595, 630 (2017) ("[A] series of such decisions will yield a refined principle or rule, resulting in fewer injustices and inefficiencies than would result if the first court's approach were followed religiously in all similar cases.").

[77] See Aharon Barak, The Judge in a Democracy 156 (2006) (noting that expansive judicial case law hangs on narrow statutory hooks and that judges develop common law within the frameworks of statutes).

[78] Cuéllar, supra note 54, at 154–56 (discussing varying levels of deference depending on the seniority of an agency decisionmaker); id. at 158 (asking whether courts should revisit the presumption of regularity to "ensure that decisionmakers recognize the risks of relying on automated analytical techniques they do not entirely understand").

[79] David A. Strauss, Common Law Constitutional Interpretation, 63 U. Chi. L. Rev. 877, 895 (1996) ("Everyone recognizes that law . . . is in substantial part about following precedent and otherwise maintaining continuity with the past.").

[80] Many describe these tools as employing machine learning, though the companies developing the algorithms often invoke "trade secrets," which prevents both defendants and judges from knowing precisely how the algorithms function. See, e.g., Ric Simmons, Quantifying Criminal Procedure: How to Unlock the Potential of Big Data in Our Criminal

Here, too, they may—and should—demand certain explanations for how those algorithms work, to ensure that the algorithms are trustworthy and fair. Defense counsel also are likely to press prosecutors and algorithm developers for explanations, which in turn may stimulate judges to do the same.

Officials in the criminal justice system often need to predict how likely a person is to commit a dangerous act.[81] In the bail context, for example, judges must assess whether individuals are likely to return to court for trial and whether they are likely to engage in criminal acts if they are not kept in detention before trial.[82] When sentencing a defendant, the judge considers in part how likely it is that the person will reoffend if released after a particular period.[83] These data-driven algorithms have the potential to help decisionmakers avoid relying on intuition and personal biases and to allow governments to reduce jail populations without affecting public safety.[84] As a result, the criminal justice system has seen a widespread and growing use of predictive algorithms in the bail, sentencing, and parole contexts.[85]

Notwithstanding their potential, these algorithms have come under intense criticism. Some critiques focus on the idea that the data on which

Justice System, 2016 Mich. St. L. Rev. 947, 997 (discussing machine learning criminal justice algorithms and noting that judges need to understand the factors that the algorithm used and the historical accuracy of the algorithm's results). Even if criminal justice algorithms currently do not use advanced machine learning, scholars have argued that they soon will. See, e.g., Richard Berk, Criminal Justice Forecasts of Risk: A Machine Learning Approach 110–11 (2012) ("Actuarial methods are changing rapidly. Forecasts increasingly exploit enormous datasets that are routinely available in real time."); Richard Berk & Jordan Hyatt, Machine Learning Forecasts of Risk to Inform Sentencing Decisions, 27 Fed. Sent'g Rep. 222, 222 (2015) (arguing that machine learning forecasting methods will produce more accurate forecasts than more traditional regression analyses).

[81] See Ashley Deeks, Predicting Enemies, 104 Va. L. Rev. 1529, 1538 (2018).

[82] Samuel R. Wiseman, Fixing Bail, 84 Geo. Wash. L. Rev. 417, 420–21 (2016) (discussing dangerousness and flight risk as two key considerations in bail decisionmaking).

[83] See Sonja B. Starr, Evidence-Based Sentencing and the Scientific Rationalization of Discrimination, 66 Stan. L. Rev. 803, 809 (2014) (noting that evidence-based sentencing is designed to assist judges in pursuit of sentencing objectives that are centered on reducing the defendant's future crime risk).

[84] Sam Corbett-Davies, Sharad Goel & Sandra González-Bailón, Even Imperfect Algorithms Can Improve the Criminal Justice System, N.Y. Times (Dec. 20, 2017), https://www.nytimes.com/2017/12/20/upshot/algorithms-bail-criminal-justice-system.html [https://perma.cc/E7BQ-EDCK].

[85] See Algorithms in the Criminal Justice System, Elec. Privacy Info. Ctr., https://epic.org/algorithmic-transparency/crim-justice/ [https://perma.cc/3BES-5P2L] (last visited Aug. 3, 2019) (listing different states' uses of algorithmic tools for sentencing, probation, and parole decisions). For a recent example of a state's decision to require the use of algorithms in the bail setting, see Dave Gershgorn, California Just Replaced Cash Bail with Algorithms, Quartz (Sept. 4, 2018), https://qz.com/1375820/california-just-replaced-cash-bail-with-algorithms/ [https://perma.cc/JRM4-RX5Z].

computer scientists train the algorithms are racially biased.[86] Others argue that the algorithms are no better at predicting recidivism than are humans who lack criminal justice expertise.[87] Finally, many object to the fact that the algorithms' structure, contents, and testing are opaque.[88] This latter concern came to a head in *State v. Loomis*, a case in which a defendant challenged the judge's use of a sentencing algorithm called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) that had categorized him as posing a "high risk of recidivism."[89] The defendant argued that the court's use of the risk assessment violated his due process rights, in part because he was not able to assess COMPAS's accuracy.[90]

The Wisconsin Supreme Court upheld his sentence.[91] Nevertheless, the majority and a concurring Justice expressed caution about the use of opaque sentencing algorithms. The majority required future presentence investigation reports to contain warnings about the limitations of COMPAS, in order to avoid potential due process violations.[92] In concurrence, Justice Shirley Abrahamson stated that "this court's lack of understanding of COMPAS was a significant problem in the instant case."[93] She noted that "making a record, including a record explaining consideration of the evidence-based tools and the limitations and strengths thereof, is part of the long-standing, basic requirement that a circuit court explain its exercise of discretion at sentencing."[94] Even the U.S. government brief, filed to oppose the defendant's petition for writ of certiorari in the U.S. Supreme Court, conceded that "[s]ome uses of an undisclosed risk-assessment algorithm might raise due process concerns—if, for example, a defendant is denied access to the factual inputs

[86] See, e.g., Andrew Guthrie Ferguson, The Rise of Big Data Policing 131–32 (2017) ("Police data remains colored by explicit and implicit bias.").

[87] See, e.g., State v. Loomis, 881 N.W.2d 749, 775 n.3 (Wis. 2016) (Abrahamson, J., concurring) (acknowledging that studies differ on the accuracy of the recidivism scores of Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)); Dressel & Farid, supra note 18, at 1 (concluding that nonexperts are "as accurate and fair as COMPAS at predicting recidivism" and noting the inefficacy of its more sophisticated features).

[88] See, e.g., Andrea Roth, Trial by Machine, 104 Geo. L.J. 1245, 1270 (2016) (noting concerns about obscuring hidden subjectivities and errors in criminal justice algorithms); Rebecca Wexler, Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System, 70 Stan. L. Rev. 1343, 1349–50 (2018) [hereinafter Wexler, Life, Liberty] ("Developers often assert that details about how their tools function are trade secrets.").

[89] *Loomis*, 881 N.W.2d at 753, 755.

[90] Id. at 757.

[91] Id. at 772.

[92] See id. at 769–70.

[93] Id. at 774 (Abrahamson, J., concurring) (noting that, "[a]t oral argument, the court repeatedly questioned both the State's and defendant's counsel about how COMPAS works" but that "[f]ew answers were available").

[94] Id. at 775.

about his criminal and personal history, or if his risk scores form part of a sentencing 'matrix' or establish a 'presumptive' term of imprisonment."[95]

Perhaps not surprisingly, some jurisdictions are shifting away from opaque commercial algorithms such as the one used in *Loomis* and toward algorithms that use public data and publicly available source codes.[96] Even those jurisdictions may retain an interest in xAI, because source codes alone are typically not self-explanatory. In particular, though, it is the courts in jurisdictions that continue to rely on opaque predictive algorithms that may—and should—become more aggressive in demanding xAI. There are a host of reasons why they might do so. First, both federal and state courts often face statutory requirements to justify the sentences they impose.[97] This allows the public to evaluate the reasonableness of the sentence and see what factual findings the judge made; it also permits review by appellate courts.[98] Judges who rely in part on sentencing algorithms might believe that they need to understand the parameters of the algorithms to articulate the reasons for using them. Second, judges serve as a bulwark to ensure accuracy and fairness in sentencing; demanding xAI will help judges evaluate whether the algorithms meet that standard or contain significant errors.[99] Third, judges might demand xAI to ensure the institutional integrity of the courts, which is undercut if courts use unreliable sources of guidance. Fourth, judges

[95] Brief for the United States as Amicus Curiae at 18, Loomis v. Wisconsin, 137 S. Ct. 2290 (2017) (No. 16-6387), 2017 WL 2333897.

[96] See, e.g., Elaine Angelino, Nicholas Larus-Stone, Daniel Alibi, Margo Seltzer & Cynthia Rudin, Learning Certifiably Optimal Rule Lists for Categorical Data, J. Machine Learning Res., June 2018, at 1, 2, http://www.jmlr.org/papers/volume18/17-716/17-716.pdf [https://perma.cc/LA46-XA6R] (developing an open-source machine learning model that accurately predicts a person's likelihood of rearrest); Creating a Fairer Pretrial System, Arnold Ventures (Dec. 1, 2017), https://www.arnoldventures.org/stories/creating-a-fairer-pretrial-system [https://perma.cc/TP87-G2L6] (stating that roughly forty jurisdictions have adopted or are in the process of implementing a pretrial risk assessment tool called the Public Safety Assessment); Public Safety Assessment: Risk Factors and Formula, Pub. Safety Assessment, https://www.psapretrial.org/about/factors [https://perma.cc/W3P3-4UJN] (last visited Aug. 3, 2019) (disclosing the risk factors and formula undergirding the Public Safety Assessment).

[97] See, e.g., 18 U.S.C. § 3553(c) (2012) (requiring that the "court, at the time of sentencing, shall state in open court the reasons for its imposition of the particular sentence," including specific reasons for sentencing outside the range); Heather Young Keagle, Appellate Div., N.J. Superior Court, Manual on New Jersey Sentencing Law 11 (rev. 2019), https://njcourts.gov/attorneys/assets/attyresources/manualsentencinglaw.pdf [https://perma.cc/UQ77-7U32] ("At the time of sentencing, the court must 'state reasons for imposing such sentence including . . . the factual basis supporting a finding of particular aggravating or mitigating factors affecting sentence.'" (quoting State v. Fuentes, 85 A.3d 923, 932 (N.J. 2014))).

[98] Toby D. Slawsky, The Importance of Statements of Reasons in Guideline Sentencing, 28 Fed. Sent'g Rep. 174, 174 (1990).

[99] See Rebecca Wexler, Code of Silence, Wash. Monthly (June/July/Aug. 2017), https://washingtonmonthly.com/magazine/junejulyaugust-2017/code-of-silence/ [https://perma.cc/5PHM-6SPV] (describing how opacity can conceal flaws in criminal justice algorithms).

might insist on some form of xAI because they are worried about being reversed on appeal for relying on a flawed or poorly understood tool. Finally, judges may demand xAI at the behest of defense counsel, to facilitate adversarial challenges and promote procedural fairness.

What form is xAI likely to take here? In the administrative law context, the audiences for the xAI (executive agencies, judges, and corporate or interest-group plaintiffs) are likely to be sophisticated actors. In the criminal justice setting, there are three main audiences: (1) judges, (2) defendants, and (3) their lawyers. Some judges and defense counsel will be sophisticated repeat players, but the defendants themselves are likely to have little experience with algorithms—and indeed judges themselves will have different levels of experience with tools such as regression analyses.[100] Judges might be more interested in model-centric explanations, while recognizing that defendants may need subject-centric xAI. Both audiences might benefit from being able to run counterfactuals through the system as well. Judges will have to decide whether to demand one or the other forms of xAI—or both.

Judges will encounter pushback from the producers of proprietary algorithms, who have resisted revealing information about the workings of their algorithms on the basis of trade secrets claims.[101] There are ways to protect such secrets, however, including by issuing protective orders.[102] Further, it might be possible to build a surrogate model of the sentencing algorithm that sheds light on its functioning without forcing the producer to reveal trade secrets. In those cases, xAI may play an important role in counterbalancing trade secrets claims such as those in play in *Loomis.*

There is another, less obvious advantage to judges' use of xAI in the criminal justice setting. A persistent concern about machine learning algorithms is that they produce "automation bias"—a tendency to unduly accept a machine's recommendation.[103] Putting xAI in front of judges

[100] This level of experience presumably will increase over time as more machine learning tools find their way into the practice of law.

[101] See Wexler, Life, Liberty, supra note 88, at 1349–50.

[102] See id. at 1409–10.

[103] Kate Goddard, Abdul Roudsari & Jeremy C. Wyatt, Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators, 19 J. Am. Med. Informatics Ass'n 121, 121 (2012); Raja Parasuraman & Dietrich H. Manzey, Complacency and Bias in Human Use of Automation: An Attentional Integration, 52 Hum. Factors 381, 397 (2010) (concluding that both expert and inexpert participants suffer from complacency and bias in their interactions with automated systems).

may lead them to question an algorithm's conclusions in a way that helps them avoid succumbing to automation bias.[104]

In light of the various benefits of xAI and a growing number of xAI tools in the toolbox, one puzzle is why courts have not already begun to insist on xAI when confronted with machine learning algorithms in criminal justice settings. Is it because the idea of xAI is nascent? Because the use of algorithms in the criminal justice context is only now starting to receive widespread scrutiny and criticism? Because of trade secrets hurdles? Or because the courts themselves currently lack the confidence to understand and use xAI?[105] It is likely a combination of all of these factors. However, as the use of machine learning and, concomitantly, xAI spreads, the courtroom is a fertile ground in which to connect xAI to real-world challenges.

Conclusion

Agency rulemaking and criminal justice are hardly the only areas of law in which courts will confront machine learning algorithms. Other possible legal contexts include product liability litigation involving self-driving cars or the internet of things,[106] litigation challenging school districts' use of algorithms for teacher evaluations,[107] malpractice litigation against doctors who rely on medical algorithms for diagnoses,[108] individual challenges to governmental decisions to freeze people's assets based on algorithmic recommendations,[109] defendants' challenges to police

[104] See Matt O'Brien & Dake Kang, AI in the Court: When Algorithms Rule on Jail Time, Phys.org (Jan. 31, 2018), https://phys.org/news/2018-01-ai-court-algorithms.html [https://perma.cc/84R8-B2X4] (discussing automation bias in judges); see also id. (quoting a Northwestern University computer scientist as arguing that judges need "boxes that give [them] answers and explanations and ask [them] if there's anything [they] want to change").

[105] Lilian Edwards & Michael Veale, Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?, 16 IEEE Security & Privacy 46, 53 (2018) [hereinafter Edwards & Veale, Enslaving the Algorithm] ("It seems quite likely that courts will be reluctant to become activists about disclosures of source code, let alone algorithmic training sets and models, until they feel more confident of their ability to comprehend and use such evidence—which may take some time.").

[106] See Ian Bogost, Can You Sue a Robocar?, Atlantic (Mar. 20, 2018), https://www.theatlantic.com/technology/archive/2018/03/can-you-sue-a-robocar/556007/ [https://perma.cc/84LK-AQ9V] (discussing the legal implications of accidents caused by self-driving cars)

[107] See Coglianese & Lehr, Governance, supra note 12, at 37–38 (discussing litigation by teachers over a school district's use of algorithms to rate teachers' performance).

[108] See Shailin Thomas, Artificial Intelligence, Medical Malpractice, and the End of Defensive Medicine, Bill of Health (Jan. 26, 2017), http://blog.petrieflom.law.harvard.edu/2017/01/26/artificial-intelligence-medical-malpractice-and-the-end-of-defensive-medicine/ [https://perma.cc/7AU6-P3QJ] (discussing the interaction between malpractice litigation and use of machine learning algorithms).

[109] See Cuéllar, supra note 54, at 144 (describing the potential use of algorithms to include decisions to freeze individuals' assets).

stops based on the use of "automated suspicion" algorithms,[110] government requests for Foreign Intelligence Surveillance Act orders based on algorithmic predictions about who is a foreign agent,[111] or challenges to algorithm-driven forensic testing.[112] These cases might implicate questions of substantive or procedural due process,[113] require "arbitrary and capricious" review, or force courts to decide whether to allow expert testimony about how a given algorithm functions.[114] Some scholars have proposed the kinds of explanations courts should seek in certain types of cases,[115] but the rubber will hit the road when the courts themselves decide what is needed. Using the tools of the common law, judges can and will productively drive the advancement and fine-tuning of xAI. When deciding xAI-related questions, courts should focus on two principles that can further public law values: maximizing xAI's ability to help identify errors and biases within the algorithm, and aligning the form of xAI in a given case with the needs of the relevant audiences.

The interest in xAI is not simply a U.S. phenomenon. The European Union's General Data Protection Regulation (GDPR), which applies to

[110] See Michael Rich, Automated Suspicion Algorithms and the Fourth Amendment, 164 U. Pa. L. Rev. 871, 875–76 (2016) ("Machine learning provides a way to go one step further and use data to identify likely criminals among the general population.").

[111] See 50 U.S.C. § 1805(a) (2012) (outlining the findings necessary for a judge to enter an ex parte order approving electronic surveillance); Jim Baker, Counterintelligence Implications of Artificial Intelligence—Part III, Lawfare (Oct. 10, 2018), https://www.lawfareblog.com/counterintelligence-implications-artificial-intelligence-part-iii [https://perma.cc/M6D3-KNW9] (discussing the use of AI in counterintelligence).

[112] See Symposium on Forensic Expert Testimony, *Daubert*, and Rule 702, 86 Fordham L. Rev. 1463, 1513–15 (2018) (presenting discussion of the application of a *Daubert*-style test to algorithm-driven DNA testing by Professor Erin Murphy of the Advisory Committee on Evidence Rules); Drew Harwell, Oregon Became a Testing Ground for Amazon's Facial-Recognition Policing. But What if Rekognition Gets It Wrong?, Wash. Post (Apr. 30, 2019), https://www.washingtonpost.com/technology/2019/04/30/amazons-facial-recognition-technology-is-supercharging-local-police/ (on file with the *Columbia Law Review*) (discussing how lawyers are preparing to litigate the admissibility of facial-recognition system evidence in court).

[113] See Coglianese & Lehr, Governance, supra note 12, at 38–43.

[114] Courts already have confronted *Daubert* issues in the face of a juvenile sentencing algorithm and algorithm-assisted discovery processes. See Moore v. Publicis Groupe, 287 F.R.D. 182, 182–84, 188–89 (S.D.N.Y. 2012) (concluding that using sophisticated algorithmic tools to search for electronically stored information is acceptable and that *Daubert* did not apply at the search stage of discovery because the documents were not yet being introduced as evidence); AI Now Inst., Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems 1, 13–14 (2018), https://ainowinstitute.org/litigatingalgorithms.pdf [https://perma.cc/Y5V4-NVRF] (noting that counsel persuaded the judge that past studies had not sufficiently validated a juvenile sentencing algorithm).

[115] See, e.g., Kroll et al., supra note 36, at 637 (describing verification methods to ensure algorithms' procedural regularity).

countries and companies in the European Union, contains provisions[116] requiring what some have termed a "right to an explanation."[117] Some scholars have interpreted the GDPR to require data controllers who make decisions about individuals based "solely on automated processing" to provide those individuals with meaningful information about the logic involved in that automated decisionmaking.[118] But it remains unclear precisely what the GDPR requires and what steps states and companies must take to meet those requirements. Other countries have enacted their own domestic "explainability" requirements. France, for instance, in its Digital Republic Act, gives individuals a right to an explanation for administrative algorithmic decisions made about those individuals.[119] That law requires the administrative decisionmaker to provide a range of information about the "degree and the mode of contribution of the algorithmic processing to the decision making," including what data were processed, what the system's parameters were, and how the algorithm weighted factors.[120] Thus, U.S. common law decisions about xAI are likely to be of interest not only to U.S. federal and state judges but to foreign judges and administrative officials as well.

Nor are courts the only government actors that must navigate the costs and benefits of xAI. Congress may demand and shape the use of xAI across industries or within government via legislation, and it may also demand the use of xAI in briefings by executive agencies, including the intelligence community. Any statute regulating the use of xAI, however, necessarily must be crafted at a high level of generality. That statute may capture the basic values that Congress wants xAI to advance, but such a statute may struggle to endure in this quickly shifting landscape. Further, the likelihood that Congress will be able to act in this space is limited, if its recent actions on complicated technology issues are any guide.[121]

[116] Regulation (EU) 2016/679 of the European Parliament and of 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), arts. 15(1)(h), 22, 2016 O.J. (L 119) 43, 46.

[117] See, e.g., Bryce Goodman & Seth Flaxman, European Union Regulations on Algorithmic Decision Making and a "Right to Explanation," AI Mag., Fall 2017, at 50.

[118] See, e.g., id. at 55. But see Sandra Wachter, Brett 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, 77, 79–90 (2017) (arguing that the GDPR implements a limited "right to be informed" rather than a "right to explanation").

[119] Edwards & Veale, Enslaving the Algorithm, supra note 105, at 48–49.

[120] Id. at 48.

[121] See Ashley Deeks, Facebook Unbound?, 105 Va. L. Rev. Online 1, 6–8 (2019), http://www.virginialawreview.org/sites/virginialawreview.org/files/01.%20Final%20Deeks.pdf [https://perma.cc/QHE5-SJ7Q] ("[Congress] has failed in its efforts to legislate on the use of encryption, election security . . . , 'hacking back,' and drone safety, and it has not tried to regulate facial-recognition software. Efforts to impose federal data-privacy laws on companies are just getting underway.").

Common law xAI thus offers real promise as we head deeper into the age of algorithms. Courts will only be able to work xAI issues at the edges, looking across legal categories to draw on xAI developments in different doctrinal areas, but that work—and the response to that work by the creators and users of machine learning algorithms—may get us where we need to be.

RULEMAKING AND INSCRUTABLE AUTOMATED DECISION TOOLS

Katherine J. Strandburg[*]

Complex machine learning models derived from personal data are increasingly used in making decisions important to peoples' lives. These automated decision tools are controversial, in part because their operation is difficult for humans to grasp or explain. While scholars and policymakers have begun grappling with these explainability concerns, the debate has focused on explanations to decision subjects. This Essay argues that explainability has equally important normative and practical ramifications for decision-system design. Automated decision tools are particularly attractive when decisionmaking responsibility is delegated *and* distributed *across multiple actors to handle large numbers of cases. Such decision systems depend on explanatory flows among those responsible for setting goals, developing decision criteria, and applying those criteria to particular cases. Inscrutable automated decision tools can disrupt all of these flows.*

This Essay focuses on explanation's role in decision-criteria development, which it analogizes to rulemaking. It analyzes whether, and how, decision tool inscrutability undermines the traditional functions of explanation in rulemaking. It concludes that providing information about the many aspects of decision tool design, function, and use that can be explained can perform many of those traditional functions. Nonetheless, the technical inscrutability of machine learning models has significant ramifications for some decision contexts. Decision tool inscrutability makes it harder, for example, to assess whether decision criteria will generalize to unusual cases or new situations and heightens communication and coordination barriers between data scientists and subject matter experts. The Essay concludes with some suggested approaches for facilitating explanatory flows for decision-system design.

INTRODUCTION

Machine learning models derived from large troves of personal data are increasingly used in making decisions important to peoples' lives.[1]

[*] Alfred Engelberg Professor of Law and Faculty Director of the Information Law Institute, New York University School of Law. I am grateful for excellent research assistance from Madeline Byrd and Thomas McBrien and for summer research funding from the Filomen D. Agostino and Max E. Greenberg Research Fund.

[1] See Max Fisher & Amanda Taub, Is the Algorithmification of the Human Experience a Good Thing?, N.Y. Times: The Interpreter (Sept. 6, 2018), https://static.nytimes.com/email-content/INT_5362.html (on file with the *Columbia Law Review*).

These tools have stirred both hopes of improving decisionmaking by avoiding human shortcomings and concerns about their potential to amplify bias and undermine important social values.[2] It is often hard for humans to grasp or explain how or why machine-learning-based models map input features to output predictions because they often combine large numbers of input features in complicated ways.[3] This inherent inscrutability[4] has drawn the attention of data scientists,[5] legal scholars,[6] policymakers,[7] and others[8] to the explainability problem.

[2] Compare Susan Wharton Gates, Vanessa Gail Perry & Peter M. Zorn, Automated Underwriting in Mortgage Lending: Good News for the Underserved?, 13 Housing Pol'y Debate 369, 370 (2002) (finding that automated underwriting systems more accurately predict mortgage default than humans and result in higher approval rates for underserved applicants), and Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig & Sendhil Mullainathan, Human Decisions and Machine Predictions, 133 Q.J. Econ. 237, 268 (2017) (showing that applying machine learning algorithms to pretrial detention decisions could reduce the jailed population by forty-two percent without an increase in crime), with Jeffrey Dastin, Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women, Reuters (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [https://perma.cc/6SA7-R35L] ("Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.").

[3] See, e.g., Finale Doshi-Velez & Mason Kortz, Accountability of AI Under the Law: The Role of Explanation 9–10 (2017), https://cyber.harvard.edu/publications/2017/11/AIExplanation [https://perma.cc/AQ5V-582E]; Jenna Burrell, How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms, Big Data & Soc'y, Jan.–June 2016, at 1, 3; Aaron M. Bornstein, Is Artificial Intelligence Permanently Inscrutable?, Nautilus (Sept. 1, 2016), http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable [https://perma.cc/B562-NCUN]; see also Info. Law Inst. at N.Y. Univ. Sch. of Law with Foster Provost, Krishna Gummadi, Anupam Datta, Enrico Bertini, Alexandra Chouldechova, Zachary Lipton & John Nay, Modes of Explanation in Machine Learning: What Is Possible and What Are the Tradeoffs?, *in* Algorithms and Explanations (Apr. 27, 2017), https://youtu.be/U0NsxZQTktk (on file with the *Columbia Law Review*).

[4] See Andrew D. Selbst & Solon Barocas, The Intuitive Appeal of Explainable Machines, 87 Fordham L. Rev. 1085, 1094 (2018) (defining "inscrutability" in this context as "a situation in which the rules that govern decision-making are so complex, numerous, and interdependent that they defy practical inspection and resist comprehension").

[5] See generally Finale Doshi-Velez & Been Kim, Towards a Rigorous Science of Interpretable Machine Learning, *in* 2018 IEEE 5th International Conference on Data Science and Advanced Analytics 1 (2018) (cataloging various ways to define and evaluate interpretability in machine learning); Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter & Lalana Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning, *in* 2018 IEEE 5th International Conference on Data Science and Advanced Analytics 80 (2018) ("While interpretability is a substantial first step, these mechanisms need to *also* be complete, with the capacity to defend their actions, provide relevant responses to questions, and be audited."); Zachary C. Lipton, The Mythos of Model Interpretability, ACMQueue (July 17, 2018), https://queue.acm.org/detail.cfm?id=3241340 [https://perma.cc/CZH3-S9JG] (discussing "the feasibility and desirability of different notions of interpretability" in machine learning).

[6] See, e.g., 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, 19–22 (2017); Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten,

This discourse has focused primarily on explanations provided to decision subjects. For example, the European Union's General Data Protection Regulation (GDPR) arguably gives decision subjects a "right to explanation,"[9] reflecting the common premise that "[t]o justify a decision-making procedure that involves or is constituted by a machine learning model, *an individual subject to that decision-making procedure* requires an explanation of how the machine learning model works."[10] Some scholars have criticized this focus, emphasizing the importance of public

Joel R. Reidenberg, David G. Robinson & Harlan Yu, Accountable Algorithms, 165 U. Pa. L. Rev. 633, 636–42 (2017); Selbst & Barocas, supra note 4; Andrew D. Selbst, Response, A Mild Defense of Our New Machine Overlords, 70 Vand. L. Rev. En Banc 87, 88–89 (2017), https://cdn.vanderbilt.edu/vu-wp0/wp content/uploads/sites/278/2017/05/23184939/A-Mild-Defense-of-Our-New-Machine-Overlords.pdf [https://perma.cc/MCW7-X89L]; Tal Z. Zarsky, Transparent Predictions, 2013 U. Ill. L. Rev. 1503, 1506–09; Robert H. Sloan & Richard Warner, When Is an Algorithm Transparent?: Predictive Analytics, Privacy, and Public Policy, IEEE Security & Privacy, May/June 2018, at 18, 18.

[7] See, e.g., Algorithmic Accountability Act of 2019, S. 1108, 116th Cong. (2019).

[8] See, e.g., Reuben Binns, Algorithmic Accountability and Public Reason, 31 Phil. & Tech. 543, 543–45 (2018); Tim Miller, Explanation in Artificial Intelligence: Insights from the Social Sciences, 267 Artificial Intelligence 1, 1–2 (2019); Brent Mittelstadt, Chris Russell & Sandra Wachter, Explaining Explanations in AI, *in* FAT*'19 at 279, 279 (2019); Deirdre K. Mulligan, Daniel N. Kluttz & Nitin Kohli, Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions, *in* After the Digital Tornado (Kevin Werbach ed., forthcoming 2020) (manuscript at 1–2), https://ssrn.com/abstract=3311894 (on file with the *Columbia Law Review*); Sandra Wachter, Brent Mittelstadt & Chris Russell, Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, 31 Harv. J.L. & Tech. 841, 842–44 (2018); Brent Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter & Luciano Floridi, The Ethics of Algorithms: Mapping the Debate, Big Data & Soc'y, July–Dec. 2016.

[9] The GDPR requires that data subjects be informed of "the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) 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." Commission Regulation 2016/679, art. 13(2)(f), 2016 O.J. (L 119) 1.

It further provides a limited "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." Id. art. 22(1). For the debate about what the GDPR's requirements entail, see, e.g., Bryan Casey, Ashkon Farhangi & Roland Vogl, Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise, 34 Berkeley Tech. L.J. 143, 153–68 (2019); Talia B. Gillis & Josh Simons, Explanation < Justification: GDPR and the Perils of Privacy, Pa. J.L. & Innovation (forthcoming 2019) (manuscript at 2–4), https://ssrn.com/abstract=3374668 (on file with the *Columbia Law Review*); Margot E. Kaminski, The Right to an Explanation, Explained, 34 Berkeley. Tech. L.J. 189, 192–93 (2019); Andrew D. Selbst & Julia Powles, Meaningful Information and the Right to Explanation, 7 Int'l Data Privacy L. 233, 233–34 (2017); Michael Veale & Lilian Edwards, Clarity, Surprises, and Further Questions in the Article 29 Working Part Draft Guidance on Automated Decision-Making and Profiling, 34 Computer L. & Security Rev. 398, 398–99 (2018); Wachter et al., supra note 8, at 861–65; Andy Crabtree, Lachlan Urquhart & Jiahong Chen, Right to an Explanation Considered Harmful (Apr. 8, 2019) (unpublished manuscript), https://ssrn.com/abstract=3384790 (on file with the *Columbia Law Review*).

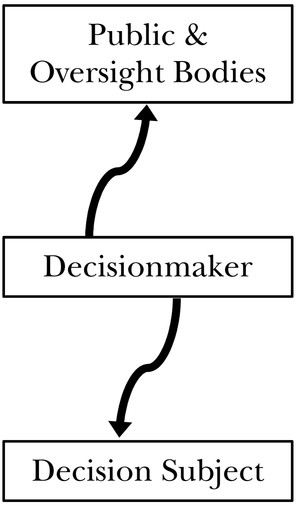
[10] Gillis & Simons, supra note 9 (manuscript at 11) (emphasis added).

accountability.[11] Talia Gillis and Josh Simons, for example, contrast "[t]he focus on individual, technical explanation . . . driven by an uncritical bent towards transparency" with their argument that "[i]nstitutions should justify their choices about the design and integration of machine learning models not to individuals, but to empowered regulators or other forms of public oversight bodies."[12] Taken together, these threads suggest the view of explanatory flows in decisionmaking illustrated in Figure 1, in which decisionmakers justify their choices by explaining case-by-case outcomes to decision subjects and separately explaining design choices regarding automated decision tools to the public and oversight bodies.

[11] For the most part, this emphasis is recent. See, e.g., Doshi-Velez & Kortz, supra note 3, at 3–9 (describing the explanation system's role in public accountability); Hannah Bloch-Wehba, Access to Algorithms, 88 Fordham L. Rev. (forthcoming 2019) (manuscript at 4–9), https://ssrn.com/abstract=3355776 (on file with the *Columbia Law Review*) ("These features . . . have prompted calls for new mechanisms of transparency and accountability in the age of algorithms."); Robert Brauneis & Ellen P. Goodman, Algorithmic Transparency for the Smart City, 20 Yale J.L. & Tech. 103, 132 (2018) ("Such accountability requires not *perfect* transparency . . . but . . . *meaningful* transparency."); Gillis & Simons, supra note 9 (manuscript at 11–12) ("Explanations of machine learning models are certainly not sufficient for many of the most important forms of justification in modern democracies . . . ."); Selbst & Barocas, supra note 4, at 1087 ("[F]aced with a world increasingly dominated by automated decision-making, advocates, policymakers, and legal scholars would call for machines that can explain themselves."); Jennifer Cobbe, Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making, Legal Stud. (July 9, 2019), https://www.cambridge.org/core/journals/legal-studies/article/administrative-law-and-the-machines-of-government-judicial-review-of-automated-publicsector-decisionmaking/09CD6B470DE4ADCE3EE8C94B33F46FCD/core-reader (on file with the *Columbia Law Review*) ("Legal standards and review mechanisms which are primarily concerned with decision-making processes, which examine how decisions were made, cannot easily be applied to opaque, algorithmically-produced decisions."). But, for a truly pathbreaking consideration of these issues, see Danielle Keats Citron, Technological Due Process, 85 Wash. U. L. Rev. 1249, 1258 (2008) ("This technological due process provides new mechanisms to replace the procedural regimes that automation endangers.").

[12] Gillis & Simons, supra note 9 (manuscript at 6–12); see also David Lehr & Paul Ohm, Playing with the Data: What Legal Scholars Should Learn About Machine Learning, 51 U.C. Davis L. Rev. 653, 708–09 (2017) (emphasizing the many choices involved in implementing a machine learning model and the different sorts of explanations that could be made).

FIGURE 1: SCHEMATIC OF EXPLANATORY FLOWS IN A SIMPLE DECISION SYSTEM

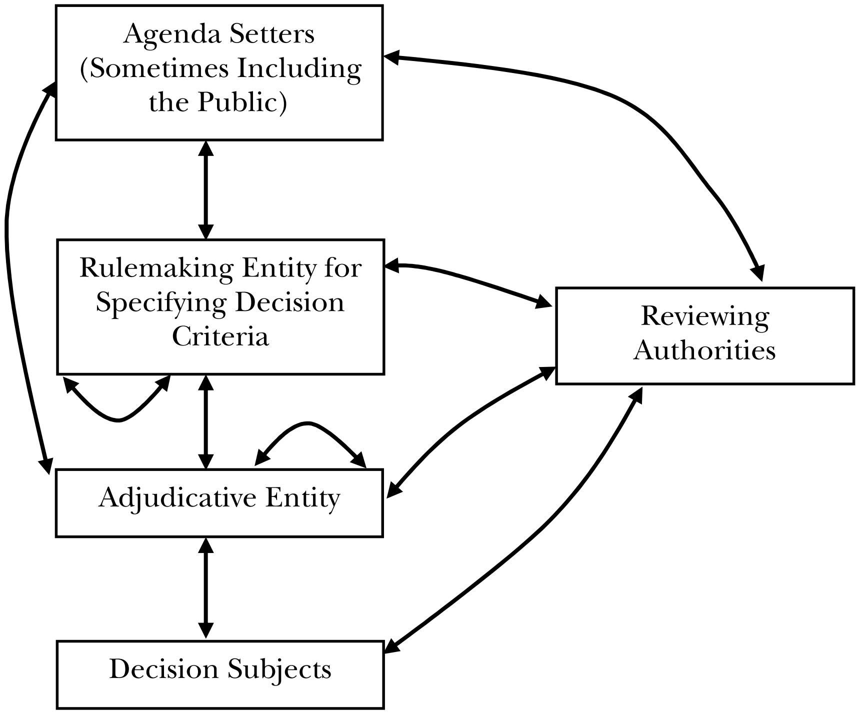


Many real-world decision systems require significantly more complex explanatory flows, however, because decisionmaking responsibility is *delegated* and *distributed* across multiple actors to handle large numbers of cases. Delegated, distributed decision systems commonly include agenda setters, who determine the goals and purposes of the systems; rulemakers tasked with translating agenda setters' goals into decision criteria; and adjudicators, who apply those criteria to particular cases.[13] In democracies, the ultimate agenda setter for government decisionmaking is the public, often represented by legislatures and courts. The public also has a role in agenda setting for many private decision systems, such as those related to employment and credit.[14] Figure 2 illustrates the explanatory flows required by a delegated, distributed decision system.

[13] The terms "adjudication" and "rulemaking" are borrowed, loosely, from administrative law. See 5 U.S.C. § 551 (2012); see also, e.g., id. §§ 553–557. The general paradigm in Figure 2 also describes many private decision systems.

[14] See infra section III.B.2.

FIGURE 2: SCHEMATIC OF EXPLANATORY FLOWS IN A DELEGATED, DISTRIBUTED DECISION SYSTEM


Agenda Setters (Sometimes Including the Public)
Rulemaking Entity for Specifying Decision Criteria
Reviewing Authorities
Adjudicative Entity
Decision Subjects

Delegation and distribution of decisionmaking authority, while often necessary and effective for dealing with agenda setters' limited time and expertise, proliferate explanatory information flows. *Delegation*, whether from the public or a private agenda setter, creates the potential for principal–agent problems and hence the need for accountability mechanisms.[15] Explanation requirements, including a duty to inform principals of facts that "the principal would wish to have" or "are material to the agent's duties," are basic mechanisms for ensuring that agents are accountable to principals.[16] *Distribution* of responsibility multiplies these principal–agent concerns, while adding an underappreciated layer of

[15] See Kathleen M. Eisenhardt, Agency Theory: An Assessment and Review, 14 Acad. Mgmt. Rev. 57, 61 (1989) ("The agency problem arises because (a) the principal and the agent have different goals and (b) the principal cannot determine if the agent has behaved appropriately."); see also Gillis & Simons, supra note 9 (manuscript at 6–10) (arguing for a principal–agent framework of accountability in considering government use of machine learning).

[16] Restatement (Third) of Agency § 8.11 (Am. Law Inst. 2005).

explanatory flows necessary for coordination among decision-system actors.[17]

Automated decision tools are particularly attractive to designers of delegated, distributed decision systems because their deployment promises to improve consistency, decrease bias, and lower costs.[18] For example, such tools are being used or considered for decisions involving pretrial detention,[19] sentencing,[20] child welfare,[21] credit,[22] employment,[23] and tax auditing.[24] Unfortunately, the inscrutability of many machine-learning-based decision tools creates barriers to all of the explanatory flows illustrated in Figure 2.[25] Expanding the focus of the explainability debate to include public accountability is thus only one step toward a more realistic view of the ramifications of decision tool inscrutability. Before incorporating machine-learning-based decision tools into a delegated, distributed decision system, agenda setters should have a clear-eyed view of what information is feasibly available to all of the system's actors. This would enable them to assess whether that information, combined with other mechanisms, can provide a sufficient level of accountability[26] and coordination to justify the use of a particular automated decision tool in a particular context.

[17] See supra Figure 2.

[18] See, e.g., Cary Coglianese & David Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era, 105 Geo. L.J. 1147, 1160 (2017) [hereinafter Coglianese & Lehr, Regulating by Robot] ("Despite this interpretive limitation, machine-learning algorithms have been implemented widely in private-sector settings. Companies desire the savings in costs and efficiency gleaned from these techniques . . . .").

[19] See, e.g., Jessica M. Eaglin, Constructing Recidivism Risk, 67 Emory L.J. 59, 61 (2017).

[20] See, e.g., State v. Loomis, 881 N.W.2d 749, 753 (Wis. 2016).

[21] See, e.g., Dan Hurley, Can an Algorithm Tell When Kids Are in Danger?, N.Y. Times Mag. (Jan. 2, 2018), https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html (on file with *the Columbia Law Review*).

[22] See, e.g., Matthew Adam Bruckner, The Promise and Perils of Algorithmic Lenders' Use of Big Data, 93 Chi.-Kent L. Rev. 3, 12–13 (2018).

[23] See, e.g., Pauline T. Kim, Data-Driven Discrimination at Work, 58 Wm. & Mary L. Rev. 857, 860 (2017).

[24] See, e.g., Kimberly A. Houser & Debra Sanders, The Use of Big Data Analytics by the IRS: Efficient Solutions or the End of Privacy as We Know It?, 19 Vand. J. Ent. & Tech. L. 817, 819–20 (2017).

[25] See infra section IV.B.

[26] See, e.g., Bloch-Wehba, supra note 11 (manuscript at 27–28) (discussing the challenge of determining adequate public disclosure of algorithm-based government decision-making); Brauneis & Goodman, supra note 11, at 166–67 ("Governments should consciously generate—or demand that their vendors generate—records that will further public understanding of algorithmic processes."); Citron, supra note 11, at 1305–06 (arguing that mandatory audit trails "would ensure that agencies uniformly provide detailed notice to individuals"); Gillis & Simons, supra note 9 (manuscript at 2) ("Accountability is achieved when an institution must justify its choices about how it developed and implemented its decision-making procedure, including the use of statistical techniques or machine learning, to an individual or institution with meaningful powers of oversight and

Incorporating inscrutable automated decision tools has ramifications for all stages of delegated, distributed decisionmaking. This Essay focuses on the implications for the creation of decision criteria—or rulemaking.[27] As background for the analysis, Part I briefly compares automated, machine-learning-based decision tools to more familiar forms of decisionmaking criteria. Part II uses the explanation requirements embedded in administrative law as a springboard to analyze the functions that explanation has conventionally been expected to perform with regard to rulemaking. Part III considers how incorporating inscrutable machine-learning-based decision tools changes the potential effectiveness of explanations for these functions. Part IV concludes by suggesting approaches that may alleviate these problems in some contexts.

I. Incorporating Machine-Learning-Based Tools into Delegated, Distributed Decision Systems

The design of a delegated, distributed decision system begins with an agenda setter (or agenda setters) empowered to determine the goals that should guide case-by-case decisions. To align decision outcomes with the system's goals as consistently and efficiently as possible, agenda setters task rulemakers with specifying decision criteria for adjudicators to apply. While legislators specify some decision criteria on behalf of the public, they routinely delegate rulemaking to agencies.[28] The general framework of agenda setting, rulemaking, and adjudication describes many decision systems, including in the private sector.[29]

A. *Rules, Standards, and Automated Decision Tools*

Rulemakers can devise various sorts of decision criteria, depending on the decision context. Criteria can be rule-like—specifying which case-by-case facts are to be taken into account and how—or standard-like—giving adjudicators more flexibility regarding what factual circumstances

enforcement."); Selbst & Barocas, supra note 4, at 1138 ("Where intuition fails, the task should be to find new ways to regulate machine learning so that it remains accountable.").

[27] Elsewhere, I focus on the implications for adjudication. Katherine J. Strandburg, Adjudicating with Inscrutable Decision Rules, *in* Machine Learning and Society: Impact, Trust, Transparency (Marcello Pelillo & Teresa Scantamburlo eds., forthcoming 2020) (on file with the *Columbia Law Review*).

[28] See Whitman v. Am. Trucking Ass'ns, 531 U.S. 457, 488 (2001) (Stevens, J., concurring in part and concurring in the judgment) ("[I]t would be both wiser and more faithful to what we have actually done in delegation cases to admit that agency rulemaking authority is 'legislative power.'"); see also, e.g., 5 U.S.C. § 553 (2012) (explaining the process by which agencies engage in informal rulemaking); 42 U.S.C. § 7409 (2012) (delegating determination of emissions and other air pollution standards to the Environmental Protection Agency).

[29] See Tony Porter & Karsten Ronit, Self-Regulation as Policy Process: The Multiple and Criss-Crossing Stages of Private Rule-Making, 39 Pol'y Sci. 41, 43 (2006) (explaining how private firms develop policy to avoid government regulation using processes such as agenda setting, problem identification, and adjudication).

they deem relevant and how they weigh those facts in coming to a decision. Decision criteria may also combine rule-like and standard-like aspects according to various schemes. For example, DWI laws in many states combine a rule-like blood alcohol threshold, above which a finding of intoxication is required, with a standard-like evaluation of intoxication at lower levels.[30] Some speed limit laws use a somewhat different scheme: Above a rule-like speed limit, there is a presumption of unsafe driving, but adjudicators may make standard-like exceptions for a narrow range of emergency circumstances.[31]

Federal sentencing guidelines illustrate another possible approach. In *United States v. Booker*, the Supreme Court held that it is unconstitutional to treat the guidelines as completely mandatory rules.[32] Judges are now "required to properly calculate and consider the guidelines when sentencing, even in an advisory guideline system."[33] The guidelines thus retain their rule-like character, but the combination scheme now gives judges the flexibility to weigh them in light of other circumstances.

Rulemakers' design choices implicate well-known trade-offs between the predictability, consistency, technical expertise, and efficiency of rule-like criteria on the one hand and the flexibility and adaptability of standard-like criteria on the other. Incorporating an automated decision tool has several implications for those design choices. First, rulemakers will need to divide decision criteria explicitly into automated and nonautomated sets, recognizing that automated assessment is utterly rule-like. Conventional narrative descriptions of decision criteria allow a spectrum from rule-like to standard-like that does not always demand such bright line allocation up front. Second, automation, especially using machine learning, distinctively constrains the sorts of rules that can be developed.[34] Third, the use of inscrutable automated decision tools limits the schemes that adjudicators can feasibly use to combine automated assessments with their assessments of nonautomated factors.[35]

Complete automation of consequential decisions is uncommon, and likely to remain so, for normative and legal reasons.[36] Human adjudicators will often be tasked with evaluating some aspects of decision criteria and combining those evaluations with automated tool outputs to make final decisions. Because different combination schemes can produce very

[30] See Drunk Driving Laws and Penalties by State, Justia, https://www.justia.com/50-state-surveys/drunk-driving-dui-dwi/ [https://perma.cc/8B35-AKUP] (last updated July 2018). For a specific example, see N.Y. Veh. & Traf. Law § 1192 (McKinney 2019).

[31] See, e.g., Speeding Tickets: How to Defend Yourself, Nolo, https://www.nolo.com/legal-encyclopedia/speeding-tickets-defending-yourself-29605.html [https://perma.cc/AQ3L-469A] (last visited Aug. 10, 2019).

[32] 543 U.S. 220, 245 (2005).

[33] U.S. Sentencing Comm'n, Guidelines Manual 14 (2018).

[34] See Strandburg, supra note 27 (manuscript at 16); infra section II.B.

[35] See Strandburg, supra note 27 (manuscript at 13–20).

[36] See infra Part III.

different outcomes, rulemakers should specify a combination scheme for adjudicators to apply. The rigidity of automated assessment rules limits the feasible combination schemes, especially when the automated tool is inscrutable to adjudicators, and often to rulemakers as well.[37]

B. *Machine Learning Models as Decision Tools*

Developments in machine learning are driving the recent upsurge of interest in automated decision tools. Machine learning is designed to fit "big" training data to complex, nonlinear models that map large sets of input features to outcome variables,[38] which serve as proxies for a relevant decision criterion of interest.[39] By using large numbers of features and training data for many individual cases, machine learning can automatically "learn" nuanced distinctions between cases from the training data, thereby producing models that are both "personalized" and more "evidence based" than may be possible using more conventional rulemaking approaches.[40] The hope is that machine-learning-based decision tools can extend automated, rule-like assessment to some decision criteria that adjudicators would conventionally have been required to evaluate in a more standard-like manner.[41] The choice to incorporate a machine-learning-based decision tool constrains rulemakers' design choices in several important ways, however.

[37] For more in-depth discussion of this point, see Strandburg, supra note 27 (manuscript at 13–19).

[38] See John Nay & Katherine J. Strandburg, Generalizability: Machine Learning and Humans in the Loop, *in* Research Handbook on Big Data Law (Roland Vogl ed., forthcoming 2019) (manuscript at 15), https://ssrn.com/abstract=3417436 (on file with the *Columbia Law Review*).

[39] For useful overviews of the machine learning process, see Lehr & Ohm, supra note 12, at 669–702; Burrell, supra note 3, at 5; Pedro Domingos, A Few Useful Things to Know About Machine Learning, Comms. ACM, Oct. 2012, at 78, 79–80. Note that data scientists usually separate the available data into "training" and "test" data sets to improve validation. Lehr & Ohm, supra note 12, at 685–88. There are a number of techniques for doing this, but this Essay will gloss over the distinction and refer to all of the data that is used to develop the model as "training" data.

[40] See Anthony J. Casey & Anthony Niblett, A Framework for the New Personalization of Law, 86 U. Chi. L. Rev. 333, 333 (2019) ("Personalized law is an old concept. The idea that the law should be tailored to better fit the relevant context to which it applies is obvious and has been around as long as the idea of law itself."); see also P'ship for Pub. Serv., Seize the Data: Using Evidence to Transform How Public Agencies Do Business 3 (2019), https://ourpublicservice.org/wp-content/uploads/2019/06/Seize-the-Data.pdf [https://perma.cc/2UD3-D2QA] (discussing the ways in which federal agencies can utilize data to inform their decisionmaking).

[41] Casey & Niblett, supra note 40, at 335 ("As technologies associated with big data, prediction algorithms, and instantaneous communication reduce the costs of discovering and communicating the relevant personal context for a law to achieve its purpose, the goal of a well-tailored, accurate, and highly contextualized law is becoming more achievable."). But see, e.g., Solon Barocas, danah boyd, Sorelle Friedler & Hanna Wallach, Editorial, Social and Technical Trade-Offs in Data Science, 5 Big Data 71 (2017) (providing an overview of several critiques of machine learning models).

1. *Data-Driven Constraints on Rule Design.* — Machine learning has the potential to create nuanced models of how outcome variables depend on many feature variables, but collecting the sort of "big data" needed to take advantage of machine learning's strengths is difficult and expensive. As a result, machine learning processes often rely on "found data,"[42] collected for some other purpose, to train the models.[43] Unfortunately, reliance on found data leaves rulemakers at the mercy of whatever feature sets and outcome variables happen to have been collected.[44] Having "big data" for an outcome variable makes it possible to train a model that effectively predicts that outcome variable, but that sort of data is often not available for the decision criteria that are truly of interest. Treating a loose or inaccurate proxy as if it were a true assessment is likely to lead to inaccurate, biased, and otherwise problematic decisions.[45] For example, a judge might like to know the likelihood that the defendant would commit a serious crime if released pending trial, but the available data might instead record arrests for any crime, which is a loose and biased proxy for the factor of interest.[46] There is thus often a trade-off between using an outcome variable for which "bigger" data is available and using a better proxy for the true criteria of interest. The need for "big" training data similarly limits the available feature sets to data types that have been recorded for large numbers of individuals.[47] Those limits constrain the sorts of factual "evidence" that can be considered by a machine-learning-based decision tool. As a result, opting to use a machine-learning-based decision tool places restrictions on decision-criteria design that may or may not be worth the trade-offs.

The limitations imposed by training data availability are related to a machine learning model's "generalizability," or ability to perform well in handling cases that were not included in the data used to train it.[48] Generalizability is also related to issues of over- or under-fitting that are associated with the extent to which a model can pick up normatively

[42] See Matthew Salganik, Bit by Bit: Social Research in the Digital Age, ch. 2.2 (open review ed. 2019).

[43] Id.

[44] Id.

[45] Id.; see also Emily Keddell, Substantiation Decision-Making and Risk Prediction in Child Protection Systems, 12 Pol'y Q. 46, 48 (2016) (discussing bias and other problems with using "substantiation, meaning a decision that abuse has been investigated and found to have occurred," as an outcome variable for predicting risk of child abuse).

[46] See, e.g., Eaglin, supra note 19, at 75–77 (2017) ("[D]efining recidivism is less intuitive and more subjective than it may appear.").

[47] See Nay & Strandburg, supra note 38 (manuscript at 14) ("Relevant information may be left out of the feature set simply because it was not prevalent enough in the training data, because it is idiosyncratic, unquantifiable or otherwise not collectible *en masse* or because it is newly available and/or newly relevant due to societal or technological changes.").

[48] Id. (manuscript at 7) ("A model is generalizable to the extent it applies, and performs similarly well, beyond the particular dataset from which it was derived.").

relevant distinctions between cases.[49] A model that accurately fits its training data can fail to generalize well if new factual scenarios crop up over time, if its outcome variable is a bad proxy for some subgroups of the population, if its feature variables do not capture all normatively relevant distinctions, or if it is simply over-fitted to the training data because of the way that developers have tuned the machine learning parameters. By limiting the outcome variables and features that a machine-learning-based model can consider, data availability constraints are likely to limit the model's generalizability. There is no computational metric for generalizability because it depends on how well the model will perform on as-yet-unknown cases.

2. *Inscrutability and Decision-Criteria Design.* — Machine learning's inscrutability stems from the fact that the computational mapping from feature inputs to outcome prediction is often hard to explain in terms that are intuitively comprehensible to humans.[50] Part of what makes these mappings difficult to explain is their reliance on large numbers of features, which can make the behavior of even simple functions difficult to intuit.[51] "Deep learning" models lack explainability at a more fundamental level, in that the ways they map input features to outcome variables cannot be represented in standard forms, such as closed equations, decision trees, or graphs.[52] Even developers and subject matter experts find it difficult or impossible to interpret such models, though there is ongoing research into technical methods for producing approximate interpretations of inscrutable machine-learning models and for training sufficiently accurate explainable models.[53] Developers employ inscrutable machine learning models, despite their explainability issues, because they are often more accurate in fitting the training data.[54] Essentially, this is because a more complicated, and thus less explainable, computational mapping can always be fit more closely to the training data.[55] Discussions of this trade-off between "accuracy" and explainability have focused rather myopically on explanation's value to decision

[49] Id. For the balance of this Essay, I will refer to both sorts of concerns as "generalizability."

[50] See supra note 3 and accompanying text.

[51] See, e.g., Selbst & Barocas, supra note 4, at 1100–05 (discussing explainability in credit scoring).

[52] See, e.g., Bornstein, supra note 3.

[53] See, e.g., Lehr & Ohm, supra note 12, at 708–10; Selbst & Barocas, supra note 4, at 1110–15.

[54] See, e.g., David Weinberger, Optimization over Explanation: Maximizing the Benefits of Machine Learning Without Sacrificing Its Intelligence, Medium (Jan. 28, 2018), https://medium.com/berkman-klein-center/optimization-over-explanation-41ecb135763d [https://perma.cc/4U2F-5BW7].

[55] See id. ("[U]nderstanding and measuring AI systems in terms of their optimizations gives us a way to benefit from them even though they are imperfect and even when we cannot explain their particular outcomes.").

subjects.[56] This section briefly explores how inscrutability constrains decision-criteria design, focusing on its implications for a decision system's ability to cope with generalizability concerns. Explanations of decision criteria also have important functions associated with accountability and coordination, which are analyzed in Part II, below.

Generalizability is essentially the technical version of the long-standing concern that rule-like decision criteria will be insufficiently flexible and forward-thinking to produce good outcomes in real-world decisions. While machine-learning-based models can be more nuanced than conventional rules in taking account of many known features, they cannot avoid the limitations of their training data.[57]

Conventional decision systems cope with generalizability concerns in two ways. First, rulemakers can scrutinize the rules in advance and try to imagine how things might go wrong, so that the rules can be redesigned to avoid problems that would otherwise crop up in real-world cases.[58] This option is not available for inscrutable machine-learning-based models. While rulemakers can and should scrutinize the training data, features, outcome variables, and validation metrics, those methods are not equivalent to scrutinizing the logic of the rule.

Second, conventional rulemakers often provide adjudicators with some standard-like flexibility to use analogy, common sense, normative judgment, and so forth to cope with case-by-case circumstances that are not adequately treated by rule-like criteria. Human adjudicators' ability to generalize in this way is limited when they are faced with the output of an inscrutable automated decision tool because they cannot discern whether and how the tool has failed to consider relevant factual circumstances. These limitations on adjudicators' capacity to generalize constrain the sorts of schemes that rulemakers can design for combining automated and nonautomated factors. For example, while adjudicators can apply the per se blood alcohol limit discussed earlier without understanding its basis, they cannot sensibly consider whether to deviate from the sentencing guidelines in a particular case without understanding the basis for the suggested sentence.[59]

II. Conventional Reasons for Explaining Rulemaking

When critics talk about the inscrutability of machine-learning-based decision tools, a common rejoinder is that human decisionmakers are

[56] See Selbst & Barocas, supra note 4, at 1111.

[57] See Nay & Strandburg, supra note 38 (manuscript at 6).

[58] Id. (manuscript at 14–15).

[59] For a more extensive discussion of these issues, see Strandburg, supra note 27 (manuscript at 15–17).

also "black boxes,"[60] in the sense that it is impossible to know what went on in a human decisionmaker's mind before coming to a decision.[61] This rejoinder misses the mark. Reason giving is a core requirement in conventional decision systems precisely *because* human decisionmakers are inscrutable and prone to bias and error, not because of any expectation that they will, or even can, provide accurate and detailed descriptions of their thought processes. This point sharpens when one shifts from Figure 1's decisionmaking paradigm to the more realistic paradigm of Figure 2. When the decisionmaker is a distributed, multi-actor institution, explanation requirements cannot be aimed at uncovering what went on in "the" decisionmaker's mind.

Generations of legal scholars have considered the functions that explanation and reason giving can perform in delegated, distributed decision systems. Machine-learning-based decision tools ease some of the familiar challenges posed by human black boxes and create some new ones.[62] Before focusing on what is distinctive about these tools, it makes sense to learn from our experience with legal explanation requirements for human decision systems.[63] Section II.A therefore provides a brief overview of some of the primary sources for legal explanation requirements. Section II.B then discusses the primary theoretical rationales behind these requirements, as relevant to rulemakers.

A. *Legal Reason-Giving Requirements*

The principle that government decisions should be justified by reasons is well enshrined in the law, not only in the United States but also in

[60] See, e.g., Frank Pasquale, The Black Box Society: The Secret Algorithms that Control Money and Information 3–8 (2015) ("The term 'black box' is a useful metaphor . . . it can mean a system whose workings are mysterious; we can observe its inputs and outputs, but we cannot tell how one becomes the other.").

[61] See, e.g., Yavar Bathaee, The Artificial Intelligence Black Box and the Failure of Intent and Causation, 31 Harv. J.L. & Tech. 889, 891–92 & nn.11–12 (2018). See generally Cynthia Rudin, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 1 Nature Machine Intelligence 206, 208–10 (2019) (arguing that explanations of black box models "often do not make sense or do not provide enough detail to understand what the black box is doing").

[62] See infra Part IV.

[63] My aim here is not to delve into when, or whether, the law *requires* explanations for government decisions using automated tools, though that question is of obvious importance. See, e.g., Danielle Keats Citron & Frank Pasquale, The Scored Society: Due Process for Automated Predictions, 89 Wash. L. Rev. 1, 8 (2014); Citron, supra note 11, at 1301–13; Kate Crawford & Jason Schultz, Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms, 55 B.C. L. Rev. 93, 121–28 (2014); Daniel J. Steinbock, Data Matching, Data Mining, and Due Process, 40 Ga. L. Rev. 1, 64–81 (2005); Cary Coglianese & David Lehr, Adjudicating by Algorithm, Regulating by Robot, The Regulatory Review (May 22, 2017), https://www.theregreview.org/2017/05/22/coglianese-lehr-adjudicating-algorithm-regulating-robot/ [https://perma.cc/7R9L-AGDR].

other democracies.[64] Under U.S. law, reason giving is a key component of the constitutional requirement that no one be deprived by government of life, liberty, or property without "due process of law."[65] Though government decisionmakers are not always required to give reasons for their decisions, reason giving is the least common denominator of due process requirements.[66]

Administrative law is especially concerned with delegated, distributed decision systems and has been described as "the progressive submission of power to reason."[67] Where agencies engage in rulemaking, explanations address not only the individual right to due process but also concerns about separation of powers and delegation of legislative power.[68] The Administrative Procedure Act (APA), along with the Constitution's due process requirement, imposes general procedural structures and constraints that apply to most federal agencies.[69] Its purposes include informing the public about the agency's activities and providing for public participation in the rulemaking process.[70] Its provisions thus exemplify the sort of explanation requirements that law imposes on rulemakers.

Explanation and justification are at the heart of notice and comment rulemaking, the most common process by which administrative agencies promulgate regulations.[71] This dialogue with the public, who are the ultimate agenda setters for government decision systems, illustrates one aspect of explanation's function within a distributed decision system. After designing a set of regulations, an agency ordinarily must

[64] See, e.g., Jerry L. Mashaw, Reasoned Administration: The European Union, the United States, and the Project of Democratic Governance, 76 Geo. Wash. L. Rev. 99, 101 (2007) [hereinafter Mashaw, Reasoned Administration].

[65] U.S. Const. amends. V, XIV.

[66] See SEC v. Chenery Corp., 318 U.S. 80, 94 (1943) (holding that an administrative agency must provide an understandable reason for its action so that a court may review it); Martin Shapiro, The Giving Reasons Requirement, 1992 U. Chi. Legal F. 179, 197 (explaining that the European Economic Community Treaty's reason-giving requirement is roughly equivalent to due process requirements in the U.S. Constitution).

[67] Jerry L. Mashaw, Small Things Like Reasons Are Put in a Jar: Reason and Legitimacy in the Administrative State, 70 Fordham L. Rev. 17, 26 (2001) [hereinafter Mashaw, Small Things].

[68] See Kevin M. Stack, The Constitutional Foundations of *Chenery*, 116 Yale L.J. 952, 1020–21 (2007) (explaining that *Chenery* "enforces a presumption" that "require[s] Congress to condition the grant of authority to an agency on the agency's expressly stating its grounds for acting"); see also, e.g., Mashaw, Small Things, supra note 67, at 22–23.

[69] See Administrative Procedure Act, 5 U.S.C. §§ 551–557 (2012); see also U.S. Const. amends. V, XIV.

[70] U.S. Dep't of Justice, Attorney General's Manual on the Administrative Procedure Act 9 (1947).

[71] More formal rulemaking processes are required in some situations. 5 U.S.C. §§ 553(c), 556–557. Agencies may also promulgate internal procedural rules, interpretive guidance, and general policy statements without engaging in notice and comment procedures. Id. § 553(b)(A).

publish them in the Federal Register, along with a section that "discusses the merits of the proposed solution, cites important data and other information used to develop the action, and details its choices and reasoning. The agency must also identify the legal authority for issuing the rule."[72] After publication, the public is given an opportunity to comment on the proposal.[73] The agency must then consider the comments when it finalizes the rule.[74] Final rules must be published along with a statement that "sets out the goals or problems the rule addresses, describes the facts and data the agency relies on, responds to major criticisms in the proposed rule comments, and explains why the agency did not choose other alternatives."[75] Rulemakers are also required to explain any later changes to existing regulations,[76] which helps ensure that reforms are made carefully and for appropriate reasons.

Judicial oversight is another mechanism for ensuring that rules are in accord with the agenda setter's goals.[77] The record of the rulemaking process is an important basis for judicial review. Courts generally must defer to agency legal interpretation and expertise wherever the governing statute is silent or ambiguous.[78] Nonetheless, a regulation may be overturned if a reviewing court determines that it is unconstitutional; inconsistent with the governing statutory authority; or arbitrary, capricious, or an abuse of discretion.[79]

Because judges often lack the substantive expertise that would be required for effective substantive review of agency rulemaking, courts perform a so-called "hard look" review of the rulemaking record to test whether an agency approached a given rulemaking task diligently, rationally, and without pursuing conflicting agendas. Under the hard look approach to the arbitrary and capricious standard, an agency must "demonstrate that it engaged in reasoned decisionmaking by providing an

[72] A Guide to the Rulemaking Process, Office of the Fed. Register, https://www.federalregister.gov/uploads/2011/01/the_rulemaking_process.pdf [https://perma.cc/5GMH-RUZP] [hereinafter Guide to Rulemaking] (last visited Aug. 11, 2019).

[73] 5 U.S.C. § 553(c).

[74] See id.; see also United States v. Nova Scotia Food Prods. Corp., 568 F.2d 240, 252–53 (2d. Cir. 1977) (holding an agency's procedures inadequate because they ignored important considerations developed through comments).

[75] Guide to Rulemaking, supra note 72.

[76] See Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 42 (1983) (holding that, while an agency's change in policy does not need to be supported by reasons more substantial than those underpinning the original rule, it must still be explained); see also FCC v. Fox Television Stations, Inc., 556 U.S. 502, 514–15 (2009).

[77] See Chevron, U.S.A, Inc. v. NRDC, 467 U.S. 837, 842–43 (1984) (holding that courts may strike down agency regulations that clearly fall outside of the bounds that Congress set).

[78] Id. at 843.

[79] 5 U.S.C. §§ 706(2)(A)–(D).

adequate explanation for its decision,"[80] "provide the 'essential facts upon which the administrative decision was based' and explain what justifies the determination with actual evidence beyond a 'conclusory statement.'"[81] A rule will also fail the test if it "is the product of 'illogical' or inconsistent reasoning; . . . fails to consider an important factor relevant to its action, such as the policy effects of its decision or vital aspects of the problem in the issue before it; or . . . fails to consider 'less restrictive, yet easily administered' regulatory alternatives."[82]

The prospect of hard look review "on the record" gives agencies incentives to create detailed records justifying the rules they promulgate, thereby also providing incentives for agencies to make rules that *can* be justified by such records. These accountability mechanisms are far from perfect and are regularly critiqued[83] but nonetheless endure as core means for addressing the unavoidable accountability problems faced by delegated, distributed decision systems.

B. *Reasons for Explaining Rulemaking*

Legal scholars have identified many normative rationales for reason-giving requirements. While some of these rationales pertain primarily to explanations aimed at decision subjects,[84] many are relevant to this Essay's focus on rulemaking. One important category of rationales focuses on improving the *quality* of the rules, in the sense of how effectively they further the agenda setter's goals. While scholars have mostly viewed quality control through an accountability lens, the law's reason-giving requirements also facilitate coordination, as this section explains. Another category of rationales is founded in the special relationship citizens have with a democratic government, in that they are both decision subjects and agenda setters.

[80] Todd Garvey, Cong. Research Serv., R41546, A Brief Overview of Rulemaking and Judicial Review 15 (2017); see also *State Farm*, 463 U.S. at 52.

[81] Garvey, supra note 80, at 15 (quoting United States v. Dierckman, 201 F.3d 915, 926 (7th Cir. 2000)).

[82] Id. (first quoting Am. Fed'n of Gov't Emps., Local 2924 v. Fed. Labor Relations Auth., 470 F.3d 375, 380 (D.C. Cir. 2006); then quoting Cincinnati Bell Tel. Co. v. FCC, 69 F.3d 752, 761 (6th Cir. 1995)).

[83] See, e.g., Gutierrez-Brizuela v. Lynch, 834 F.3d 1142, 1152 (10th Cir. 2016) (Gorsuch, J., concurring) ("In this way, *Chevron* seems no less than a judge-made doctrine for the abdication of the judicial duty."); Stephen Breyer, Judicial Review of Questions of Law and Policy, 38 Admin. L. Rev. 363, 383 (1986) (claiming that hard look review leads to "abandonment or modification of the initial project irrespective of the merits").

[84] See Henry J. Friendly, Some Kind of Hearing, 123 U. Pa. L. Rev. 1267, 1280–81 (1975) (explaining that providing notice and grounds for the proposed action helps the individual "marshal evidence and prepare his case"); Martin H. Redish & Lawrence Marshall, Adjudicatory Independence and the Values of Due Process, 95 Yale L.J. 455, 475–91 (1986) ("The instrumental conception of due process focuses on the individual's interest in having an opportunity to convince the decisionmaker that he deserves the right at issue.").

1. *Reason Giving to Improve Quality.* — Reason giving "promotes accountability by limiting the scope of available discretion and ensuring that public officials provide public-regarding justifications for their decisions" and "facilitates transparency, which, in turn, enables citizens and other public officials to evaluate, discuss, and criticize governmental action, as well as potentially to seek legal or political reform."[85] It is also a bulwark against arbitrariness.[86] For administrative agencies, "legitimacy flows primarily from a belief in the specialized knowledge that administrative decisionmakers can bring to bear on critical policy choices. And the only evidence that this specialized knowledge has in fact been deployed lies in administrators' explanations or reasons for their actions."[87] In addition to promoting quality through accountability, reason giving might be expected to improve rule quality through the disciplining effect of "showing your work" and by facilitating communication and coordination among rulemakers. Reason giving also assists in the evaluation and reform of rules.[88]

a. *The "Show Your Work" Phenomenon.* — The "show your work" phenomenon is familiar: The very process of explaining one's reasoning is likely to improve it by highlighting loopholes, inconsistencies, and weaknesses.[89] For groups, the "show your work" phenomenon includes the benefits of deliberating to jointly produce an explanation. If rulemakers anticipate that outsiders will see, and potentially critique, their explanations, the effect is heightened, since the prospect of being exposed as sloppy, ill informed, biased, or captured should provide incentives for rulemakers to devise rules that *can* be explained and justified.

By forcing rulemakers to justify their work product in terms of appropriate goals and relevant facts, the "show your work" phenomenon may also deter bias and arbitrariness. This phenomenon will presumably

[85] Glen Staszewski, Reason-Giving and Accountability, 93 Minn. L. Rev. 1253, 1278 (2009).

[86] See Lisa Schultz Bressman, Beyond Accountability: Arbitrariness and Legitimacy in the Administrative State, 78 N.Y.U. L. Rev. 461, 473–74 (2003) (noting that the initial motivation for judicial innovations such as a "reasoned consistency" requirement for agency decisions was the prevention of arbitrariness); Christine N. Cimini, Principles of Non-Arbitrariness: Lawlessness in the Administration of Welfare, 57 Rutgers L. Rev. 451, 510–12 (2005) (arguing that accountability for and reviewability of agency decisions serve to prevent arbitrary decisionmaking).

[87] Mashaw, Reasoned Administration, supra note 64, at 117. But see Jodi L. Short, The Political Turn in American Administrative Law: Power, Rationality, and Reasons, 61 Duke L.J. 1811, 1814 (2012) (discussing the "gathering movement to reconceptualize the legitimacy of administrative agencies in terms of their political—and specifically, their presidential—accountability as opposed to their expertise, their fidelity to statutory commands, or their role as fora for robust citizen participation and deliberation" (footnotes omitted)).

[88] See Strandburg, supra note 27 (manuscript at 12–13).

[89] See In Re Expulsion of N.Y.B., 750 N.W.2d 318, 326 (Minn. Ct. App. 2008) (drawing an analogy between procedural requirements in administrative law and the "show your work" method of teaching mathematics).

be most effective when it creates self-awareness of unintentional bias and arbitrariness. But, as one commentator colorfully put it, "hypocrisy has a civilising force" in human decisionmaking.[90] Explanations facilitate scrutiny, making it more difficult to mask intentional bias.

b. *Explaining to Agenda Setters.* — Notice and comment, review by the Office of Information and Regulatory Affairs (OIRA), and judicial review exemplify the interplay of explanation and feedback between agenda setters and rulemakers. Explanations to agenda setters perform two main functions related directly to the principal–agent problems mentioned earlier.[91] The first function is accountability, which entails keeping an eye out for ways in which a rulemaking entity's bias, conflicts of interest, sloppiness, or lack of zeal might have infected the rule it devised. The second function is to catch misalignments between the agenda setter's goals and the rule's potential application to real-world case types that rulemakers may not have considered adequately (or at all). This second function relates to the generalizability concerns discussed in Part I.[92] For government decision systems, the general public is the ultimate agenda setter. Public feedback may also be vital to some private decision systems because of the value of engaging diverse perspectives in ferreting out problems of accountability and misalignment.

The benefits of a public explanation obviously depend on whether the public is willing and able to engage with it—a perennial problem. Notice and comment has been criticized because well-funded, concentrated interests are better equipped to understand the proposed rules and to use their influence to bend them to their own benefit.[93] The empirical picture is mixed. While many studies find little participation by individuals in notice and comment rulemaking,[94] some have found

[90] Guido Noto La Diega, Against the Dehumanisation of Decision-Making: Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information, 9 J. Intell. Prop. Info. Tech. & Electronic Comm. L. 3, 10 (2018).

[91] See supra notes 15–16 and accompanying text.

[92] See supra section I.B.1.

[93] See Susan Webb Yackee, Sweet-Talking the Fourth Branch: The Influence of Interest Group Comments on Federal Agency Rulemaking, 16 J. Pub. Admin. Res. & Theory 103, 105 (2005) ("[T]he notice and comment period is an important political arena where the bureaucracy frequently alters and adapts public policies to better match the preferences of interest group commenters.").

[94] See Cary Coglianese, Citizen Participation in Rulemaking: Past, Present, and Future, 55 Duke L.J. 943, 958 (2006) ("Most rules still garner relatively few overall comments and even fewer comments from individual citizens."); Marissa Martino Golden, Interest Groups in the Rule-Making Process: Who Participates? Whose Voices Get Heard?, 8 J. Pub. Admin. Res. & Theory 245, 253–54 (1998) ("Neither NHTSA nor the EPA received a single comment from an individual citizen on any of the eight rules that were examined. . . . Here, fully 9 percent of the comments HUD received on this rule were submitted by individual citizens."); Nina A. Mendelson, Rulemaking, Democracy, and Torrents of E-Mail, 79 Geo. Wash. L. Rev. 1343, 1357–59 & n.79 (2011) ("Although the right to . . . submit written comments in agency rulemaking extends to every member of the public, . . . actual participation in rulemaking is not well balanced.").

substantial participation in commenting on particular sorts of regulations by citizen groups or individuals submitting form letters.[95] While citizen groups are usually not heavily resourced, they can build up significant subject matter expertise, allowing them to submit meaningful feedback and criticism.[96] This is obviously not a complete answer to the power imbalance, but it counsels against underestimating the societal benefit of public explanations. Moreover, the power imbalances in the case-by-case decision systems of interest to us here are somewhat different from those in the standard interest group story, in which powerful regulated entities use notice and comment to influence agencies propounding environmental or consumer protection regulations.[97] Here, the affected parties are individuals, who do not have outsize power to influence the design of decision systems that are critical to their opportunities in vital arenas such as employment, credit, public benefits, criminal justice, and family life.[98] Moreover, these individuals are members of the public and thus agenda setters in their own right.

c. *Communication and Coordination Among Rulemakers.* — It almost goes without saying that the quality of outcomes from a delegated, distributed decision system depends on coordination and communication between the players, including within rulemaking entities and between rulemakers and adjudicators.[99] For conventional decision systems, explanation's coordinating function has received considerably less scholarly attention than its accountability function. This is not terribly surprising for two reasons. First, the narrative form of conventional rules makes their content somewhat self-explanatory to agenda setters and adjudicators, shifting the focus toward explaining why that content is justified. Second, explanation's coordinating function piggybacks on its accountability function. Requiring rulemakers to create explanations aimed at the public, courts, or other agenda setters indirectly provides incentives for the coordinated effort necessary to create those explanations, which

[95] See Mariano-Florentino Cuéllar, Rethinking Regulatory Democracy, 57 Admin. L. Rev. 411, 462 (2005) (studying three regulatory proceedings in which 72.1%, 98.6%, and 98.3% of comments, respectively, came from individual members of the public; in two of the proceedings, individual comments were almost exclusively form letters); Golden, supra note 94, at 253–55 (finding contributions by citizens' groups ranging from 0% to 16.7% of comments depending on the agency and regulation).

[96] See Cuéllar, supra note 95, at 450–51, 458–59 (finding, in a study of two regulatory proceedings, considerably higher values for "comment sophistication" in comments from public membership or public interest organizations than from individuals); Yackee, supra note 93, at 105 ("[I]nterest group comments provide a new source of information and expertise to the bureaucracy during the rulemaking process.").

[97] See Golden, supra note 95, at 255 (contrasting EPA and NHTSA rulemakings with "extremely limited participation by public interest or citizen advocacy groups" with HUD rulemakings where "commenters include citizen advocacy groups, individual citizens, and a wide range of government agencies").

[98] See supra notes 19–23 and accompanying text.

[99] For more on explanations between rulemakers and adjudicators, see Strandburg, supra note 27 (manuscript at 10–13).

in turn activates the "show your work" phenomenon.[100] Similarly, the record creation incentivized by hard look review[101] requires internal coordination, while the resulting record can facilitate further communication and coordination. In addition, and partly to ensure that the required explanations and record will pass muster, rulemaking bodies often impose procedures that amount to internal explanation requirements.[102]

2. *Reason Giving, Democracy, and Respect.* — Reason giving legitimates governmental decisionmaking in a democracy because, as one scholar puts it, "[a]uthority without reason is literally dehumanizing. It is, therefore, fundamentally at war with the promise of democracy, which is, after all, self-government."[103] Particularly in the context of rulemaking by unelected administrative agencies, reason-giving requirements ensure that members of the public are treated as citizens, rather than subjects: "[T]o be subject to administrative authority that is unreasoned is to be treated as a mere object of the law or political power, not a subject with independent rational capacities."[104]

Explanations also empower citizens in their agenda-setting role, by helping them to understand what the rules require, providing bases for individual and group opinion formation and advocacy, and helping minorities to identify rules that ignore or undermine their interests.[105] Reason giving thus "embodies, and provides the preconditions for, a deliberative democracy that seeks to achieve consensus on ways of promoting the public good that take the views of political minorities into account."[106]

These rationales do not have the same force for private-sector decisions, where decision subjects ordinarily do not have similar agenda-setting rights. But explaining the rationale behind decisionmaking criteria also comports with more general societal norms of fair and nonarbitrary treatment. Moreover, the public has an interest as citizens and individuals, both legally and ethically, in the fairness and reasonableness of

[100] See supra section II.B.1.a.

[101] See supra section II.A.

[102] See, e.g., Jennifer Nou, Intra-Agency Coordination, 129 Harv. L. Rev. 421, 436–37, 451 (2015) (explaining that "agency heads fac[ing] greater uncertainty regarding how to formulate and draft their regulations in ways that would withstand judicial challenge . . . can respond by creating structures and processes that lower the costs of internal information processing"); Thomas O. McGarity, The Internal Structure of EPA Rulemaking, Law & Contemp. Probs., Autumn 1991, at 57, 58–59, 90–94 (detailing the internal procedures and "team" approach used by the EPA to respond to "[t]he exigencies of external review and the practical necessities of bringing multiple perspectives within EPA to bear on the decisionmaking process").

[103] Mashaw, Reasoned Administration, supra note 64, at 118.

[104] Id. at 104.

[105] See, e.g., Staszewski, supra note 85, at 1278–84.

[106] Id. at 1278.

private decision systems that fundamentally affect people's lives.[107] Indeed, private decision systems do not operate in a legal vacuum but are subject to legal protections including, for example, antidiscrimination laws and protections against fraud. In addition, as a practical matter, some subjects of private-sector decision systems are also users or customers, whose market relationships to decisionmakers give them some leverage to demand explanations of the rules that govern those relationships.

III. EXPLAINING MACHINE-LEARNING-BASED DECISION TOOLS

This Part builds on Part II's brief sketch of the purposes of reason-giving requirements by considering how the limited explainability of machine-learning-based decision tools affects the functions that explanations have conventionally been expected to perform in connection with rulemaking. Section III.A begins by taking a more precise look at which aspects of a machine-learning-based decision tool are unexplainable. Section III.B then reflects on how each of the explanation functions described in Part II is affected by the incorporation of an inscrutable machine-learning-based decision tool.

A. *Cabining Machine Learning's Explainability Problems*

Machine learning's explainability problems reside in the inscrutability of a machine learning model's computational mapping of input features to outcome variables.[108] There are, however, many aspects of the development of machine-learning-based decision tools, and of the decision rules embedded in those tools, that are just as explainable as a rule in conventional narrative form. To assess the impact of inscrutable machine-learning-based decision tools, it is important to be precise about what can and cannot be explained.

1. *Explainable Components of a Machine-Learning-Based Decision Tool.* — In some respects, the touted "black box" nature of machine learning models[109] is not nearly all that it is cracked up to be. Many choices made in the process of creating an automated decision tool are not so different from choices made in more traditional rulemaking processes. Moreover, some of those choices are *embedded as components of the rules* of the ultimate decision system, just as similar choices are reflected in narrative rules, and can be explained in conventional fashion. *Explainable components* include:

- Separation of decision criteria into automated and non-automated aspects;

[107] Cf. Gillis & Simons, supra note 9 (manuscript at 8–9) (making a similar point about accountability).

[108] See supra Introduction.

[109] See supra notes 3–8 and accompanying text.

- Definitions of decision criteria to be assessed by the automated tool;
- Definitions of outcome variables to be used as proxies for decision criteria;
- Definitions of feature variables to be used as factual evidence in automated decision criteria assessments; and
- Combination schemes governing how adjudicators should combine automated assessments with other relevant information to make decisions.

Whether, under what circumstances, and to whom the law *requires* rulemakers to explain these components is outside the scope of this analysis, but there are no *technical* barriers to requiring such explanations.

2. *Explainable Rulemaking Record.* — Other important choices involved in creating a machine learning model are not reflected on the face of the ultimate automated decision rule but can be described and explained in a record of the rulemaking process.[110] Such choices include selecting training data, determining machine learning algorithms and technical parameters, devising validation protocols, and evaluating whether a model has been adequately validated to justify using it in a decision rule. All of these choices, and the reasons for them, could be included in a record of the development of a machine-learning-based decision tool. Most importantly, such a record could include information about the sources, demographics, and other characteristics of the training data sample; definitions of validation metrics; and results of validations and performance tests. This information plays much the same role as information about statistical and more specialized technical bases for rules that are routinely included in agency rulemaking records and facilitate hard look review by courts and the cost–benefit analysis required for some rules by OIRA.[111]

B. *Explanation, Decision System Quality, and Machine-Learning-Based Tools*

In light of the previous section's parsing of explainable and unexplainable aspects of machine-learning-based decision tools, this section explores how and why incorporating such tools into a decision system is likely to affect the functions of explanation,[112] identifying where the inscrutability of a machine learning model's computational mapping from input features to outcome variables is likely to create serious problems.

1. *The "Show Your Work" Phenomenon.* — The "show your work" phenomenon carries over straightforwardly to an automated decision tool's

[110] See Selbst & Barocas, supra note 4, at 1130–33, for a similar argument in terms of "documentation."

[111] See supra section II.A.

[112] See supra Part II.

explainable components and recordable information.[113] In essence, developers' design *choices* are all explainable, and the benefits of the "show your work" phenomenon will apply to those choices.[114] The full benefits of the "show your work" phenomenon may not be retained, however, for two reasons. First, the "show your work" phenomenon is effectuated primarily through self-awareness and thus depends on developers having sufficient incentives to create detailed and persuasive explanations. Unfortunately, common practices for developing automated decision tools undermine those incentives. Because many rulemaking entities do not have data scientists on staff, they outsource development or purchase off-the-shelf products.[115] Many of these outsourced machine-learning-based decision tools are burdened with confidentiality agreements that severely limit the explanations and records of development that are provided to rulemaking entities and may block public disclosure almost entirely.[116] Such secrecy undermines the "show your work" phenomenon. Second, the "show your work" phenomenon will not aid in resolving problems that developers cannot avoid through careful design choices and validation, as discussed further in section III.B.2.c, below.

2. *Explaining to Agenda Setters.* — This section considers how the functions of explanation to agenda setters depend on access to (i) the explainable components;[117] (ii) information about data selection, sources, and validation that could be available in a rulemaking record;[118] and (iii) a conventional narrative explanation of the way that the rule maps input features to outcome variables. Explanations to agenda setters serve accountability functions but can also be important for generalizability.[119]

[113] See supra section II.B.1.a.

[114] See Anupam Chander, The Racist Algorithm?, 115 Mich. L. Rev. 1023, 1028–29 (2017) ("[E]ven for programmers or companies who intend to discriminate, the process of coding itself is likely to cause programmers to shy away from actually encoding the discrimination.").

[115] See AI Now Inst., Algorithmic Accountability Policy Toolkit 7–9 (2018), https://ainowinstitute.org/aap-toolkit.pdf [https://perma.cc/K9BQ-VQG6] (providing an overview of algorithms used by governments and developed by private companies, such as a Medicaid eligibility tool built by IBM, surveillance technologies built by Palantir, and parole term software developed by Northpointe); see also Allegheny Cty. Dep't of Human Servs., Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions (2017), https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/04/Developing-Predictive-Risk-Models-package-with-cover-1-to-post-1.pdf [https://perma.cc/YNY3-RTCD] (highlighting a child welfare service's predictive risk model built through a public–private partnership).

[116] See, e.g., Bloch-Wehba, supra note 11 (manuscript at 9–26) (describing confidentiality constraints on the use of algorithms in Medicaid, education, and criminal law enforcement); Brauneis & Goodman, supra note 11, at 137–59 (describing the use of proprietary algorithms in policing, child welfare, public safety, and teacher performance determinations).

[117] See supra section III.A.1.

[118] See supra section III.A.2.

[119] See supra section II.B.1.b.

As noted earlier, the benefits of public explanation are often effectuated through advocacy groups.[120] To isolate the unique issues stemming from machine-learning-based decision tools, it is thus helpful to consider whether such tools can be satisfactorily explained to advocacy groups with significant substantive expertise and moderate resources, assuming that most other agenda setters, such as legislatures, courts, OIRA, or private businesses, will have at least the capacity of such groups.

a. *Explainable Components.* — The explainable components identified in section III.A.1 will be understandable to a public advocacy group with sufficient expertise and resources and can facilitate extremely valuable checks on the decision system's accountability and generalizability. For example, such a group might assess whether the proxy outcome variable is biased or unlikely to generalize to some sorts of cases; consider whether the use of some feature variables is normatively unacceptable or whether important features are missing from the list; or evaluate whether the amount of flexibility given to adjudicators in combining the automated tool output with other information is appropriate. These agenda setters can help to evaluate whether it is normatively appropriate to use a rule-like automated tool to evaluate certain decision criteria or whether a more flexible, standard-like approach should be required.[121] Though rulemakers presumably will also have considered this question, they may be prone to view automation's potential through rose-colored glasses for various reasons, such as a bias toward cost-cutting measures.[122]

b. *Data Sources and Validation.* — Explanations of data sources and validation in a rulemaking record are potentially useful for uncovering bias or sloppiness, detecting holes in the coverage of the sample set, and ensuring that all normatively relevant performance metrics have been examined. For example, unrepresentative training data is one important source of generalizability problems.[123] The public's diverse perspectives may give it an edge over rulemakers in identifying forms of representativeness that might matter for the decision criteria in question.[124]

The technical knowledge about data science that is required to understand this information may currently be beyond the capacity of many advocacy groups and other agenda setters.[125] Over time, however, advocacy groups, particularly the larger and better resourced among them, will undoubtedly upgrade their technical expertise by involving data scientists in their work, as advocacy groups have done in other

[120] See supra section II.B.1.b.

[121] See supra section I.A.

[122] See Coglianese & Lehr, Regulating by Robot, supra note 18, at 1160.

[123] See supra section I.B.1.

[124] See supra notes 105–106 and accompanying text.

[125] See supra notes 95, 116 and accompanying text.

technical arenas.[126] One concern is that there are so many decision systems—national, state, local, and private—incorporating machine-learning-based decision tools that it may be difficult for advocacy groups, many of which might be small and otherwise nontechnical in nature, to keep up with all of them. For the most part, though, if characteristics about the training data, results from performance tests, and other information discussed in section III.A.2 are included in the rulemaking record, they can be expected to perform the same explanation functions as the information in a more conventional rulemaking record.

c. *Inscrutability of the Computational Mapping from Input Features to Outcome Variable.* — Information about the explainable components, data sources, and validation studies may be sufficient for the accountability function of explanation to agenda setters, in part because those information sources provide access to the most important information available to the rulemaking entity itself. The inscrutability of machine learning models creates more fundamental problems, however, regarding the extent to which explanation can help detect generalizability problems and other unintentional misalignments between the decision system's purposes and the automated criteria.[127] In some respects, the generalizability of a rule is always a guessing game—nobody can be certain how any rule will perform "out in the wild" because there may be cases that neither agenda setters nor rulemakers could have anticipated.[128] Conventional rules, with their narrative format, nonetheless allow human readers to anticipate and identify some generalizability issues using logical inference, analogy, and common sense.

These reasoning methods are not applicable to inscrutable machine learning models, however. Moreover, computational validation tools and other statistical and mathematical analyses cannot provide the same sorts of insights about generalizability, which depend on a grasp of the logic of the rule. Researchers have invented various approaches for creating approximate explanations for a machine learning model's opaque mapping.[129] While many of these methods are designed to explain the specific

[126] See, e.g., Shobita Parthasarathy, Breaking the Expertise Barrier: Understanding Activist Strategies in Science and Technology Policy Domains, 37 Sci. & Pub. Pol'y 355, 358–60 (2010) (describing how breast cancer patient advocates found sympathetic experts to educate them about the technical complexities of their causes in order to advance their advocacy). Indeed, some advocacy groups are already beginning to do this. See, e.g., AI Now Inst., Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems 4–5 (2018), https://ainowinstitute.org/litigatingalgorithms.pdf [https://perma.cc/M82T-LR9H] (noting that organizers of a recent workshop examining litigation involving the government's use of algorithmic systems featured participation by relevant legal and scientific experts).

[127] See supra section II.B.1.b.

[128] Indeed, this is a primary justification for using standards rather than rules. See Strandburg, supra note 27 (manuscript at 13).

[129] See generally Lipton, supra note 5 (surveying the academic literature of techniques designed to render machine learning models interpretable).

results of individual cases,[130] some attempt to create more general approximate explanations, which might be useful for probing generalizability issues.[131] For example, a model trained to distinguish wolves from dogs in photographs worked well on its training data but failed on a larger set of photos.[132] The problem was that the training data was skewed—nearly all of the wolves were in snowy landscapes, so the model used the presence of snow to distinguish wolves from dogs.[133] Techniques for creating approximate explanations of the machine logic helped to identify that generalizability problem because, after receiving the explanations, nearly all human observers were able to recognize that "snow" played a key role in that logic.[134]

On the whole, though, it remains uncertain whether any of these technical approaches can replace human analysis of narrative rules. Though machine learning models are trained to reproduce the outputs that human beings assigned to the training data, the mappings they create are not likely to be similar to human mental models.[135] While the association of wolves with snow ran throughout the training data, tougher generalizability issues may arise from unanticipated or uncommon "edge" cases. Humans are reasonably good at reading rules and thinking about whether they are mistaken or have blind spots but are not similarly good at predicting an inscrutable machine learning model's blind spots. For example, a deep learning model trained to triage pneumonia patients performed very well on validation tests.[136] Researchers also created a less accurate, but explainable, model based on the same data.[137] Scrutiny of the explainable model identified a problem in the data: Pneumonia patients with asthma are *high* risk, but because they had routinely been treated in the ICU, their outcomes were good,

[130] See id. at 40–42.

[131] See Doshi-Velez & Kim, supra note 5, at 7 ("Global interpretability implies knowing what patterns are present in general (such as key features governing galaxy formation), while local interpretability implies knowing the reasons for a specific decision (such as why a particular loan application was rejected).").

[132] Marco Tulio Ribeiro, Sameer Singh & Garlos Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *in* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135, 1142–43 (2016).

[133] See id.

[134] See id.

[135] See, e.g., Kevin Hartnett, Machine Learning Confronts the Elephant in the Room, Quanta (Sept. 20, 2018), https://www.quantamagazine.org/machine-learning-confronts-the-elephant-in-the-room-20180920/ [https://perma.cc/V5DY-HW2Y] (explaining that neural networks may encounter difficulty with fundamental human tasks because of their inability to process confusing and incongruous information).

[136] See Rich Caruana, Paul Koch, Yin Lou, Marc Sturm, Johannes Gehrke & Noémie Elhadad, Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission, *in* Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1721, 1730 (2015).

[137] See id. at 1721–22.

fooling the model into treating them as *low* risk.[138] The data scientists and medical experts working on the project *could*, in principle, have foreseen that the data for asthma sufferers might be misleading, but they didn't. They identified the asthma problem only after scrutinizing the explainable model.[139] The problem with the asthma data presumably also affected the inscrutable machine learning model, but researchers would not have been able to detect it. Moreover, as the study authors noted, because the inscrutable machine-learning-based model was fit more tightly to the training data than the explainable model, "it was possible that the neural nets had learned other patterns that could put some kinds of patients at risk" that did not show up in the explainable version.[140] Without an intuitive window into the logic of the machine learning model, there was simply no way to tell.

In sum, while the explainable components and rulemaking record can give agenda setters a good grasp on accountability and some handle on potential generalizability problems, there is no doubt that both rulemakers and agenda setters will more effectively anticipate generalizability problems if they can simply read the rule. Whether such lingering generalizability concerns outweigh the benefits of using an inscrutable machine-learning-based tool for particular decision criteria in a particular context can only be a normative judgment. Agenda setters—including, where appropriate, the public—should have the final say on that trade-off.

3. *Communication and Coordination Among Rulemakers.* — In conventional rulemaking, explanations created for agenda setters may be sufficient to facilitate communication and coordination among rulemakers. Incorporating a machine-learning-based tool into a decision system increases the challenges of communication and coordination, however, because of the disciplinary barriers between substantive experts and data scientists. Those barriers both heighten the importance of explanation and increase its difficulty. Data scientists differ from traditional rulemakers in three respects: (i) they are tool-building specialists, rather than subject matter specialists; (ii) they often do not work for the rulemaking entity;[141] and (iii) trade secrecy claims and confidentiality agreements can constrain their interactions with substantive rulemakers.[142]

Because data scientists are information technologists, rulemakers may be tempted to view their work as a technical task akin to those as-

[138] See id.

[139] See id.

[140] Id. at 1722.

[141] See supra note 116 and accompanying text.

[142] See Brauneis & Goodman, supra note 11, at 153 ("The owners of proprietary algorithms will often require nondisclosure agreements from their public agency customers and assert trade secret protection over the algorithm and associated development and deployment processes.").

signed to an IT department.[143] Machine learning model development is deeply entangled with subject matter expertise and normative choices, however.[144] Data scientists' role is thus more like that of empirical economists, who are also technical specialists whose methods have broad application. Good economic modeling requires considerable substantive knowledge, however, which economists must access by collaborating with substantive experts or acquiring substantive expertise. Because they are highly contextual, economic models cannot simply be used off the shelf. Before porting them over to new situations, their underpinnings must be scrutinized to determine whether they can be appropriately adapted for use in those situations. Data scientists' design decisions are even more substantively fraught because the inscrutable models they create are used directly for assessing decision criteria. As a result, the substantive, normative, and policy assumptions underlying these choices have a direct impact on decision outcomes.

Though close communication and coordination between data scientists and substantive experts is critical, each group's unfamiliarity with the other's disciplinary knowledge will tend to impede it. When the development of automated decision tools is outsourced, those difficulties inevitably mount. Confidentiality agreements and trade secrecy claims keep information from rulemakers and discourage open communication, which only makes matters worse.[145] Documentation, user manuals, and training are traditional forms of explanation between software engineers and their clients.[146] While they may be sufficient for *users*, those explanatory forms are unlikely to facilitate the close communication and coordination required for *codevelopment* of decision criteria that incorporate machine-learning-based decision tools.

C. *Reason Giving, Democracy, Respect, and Machine Learning*

Some view the use of automated decision tools as inherently dehumanizing or disrespectful, at least in some contexts.[147] Here I do not adopt that view and hence consider whether the inscrutability of machine-learning-based decision tools creates problems for democratic and human values even when conventional rule-like decision criteria would

[143] See 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/2KH9-5SFB] (explaining the tendency to view data science as objective and infallible).

[144] See id. ("Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations.").

[145] See Brauneis & Goodman, supra note 11, at 153–54.

[146] See Technical Documentation in Software Development: Types, Best Practices, and Tools, AltexSoft, https://www.altexsoft.com/blog/business/technical-documentation-in-software-development-types-best-practices-and-tools/ [https://perma.cc/S92L-99AH] (last updated Mar. 26, 2019).

[147] See, e.g., Noto La Diega, supra note 90, at 10–11.

have been acceptable. Though democratic legitimacy and dignitary concerns are part of the standard reasons for requiring government decisionmakers to provide explanations,[148] complete explanations of all government decisions have never been required. Machine-learning-based decision tools can be explained in a limited sense, as just discussed. The question, then, is when those limited explanations are sufficient in light of these sorts of concerns.

The answer to this question is likely to depend on many contextual factors, including the nature of the decision and what is at stake, the justification for automating a particular aspect of the decision criteria, the way in which adjudicators are expected to use the automated output in coming to final decisions, and, crucially, the extent to which explainable aspects of the automated tool are, in fact, explained. Citizens will likely have or develop a sense of whether the limited explanations available in a particular context are sufficient in light of these legitimacy and dignitary values. This sort of evaluation is likely to incite controversy but is not terribly different from the normative assessments that currently go into determining whether rules are appropriately employed in various contexts or what level of "due process" is appropriate for a particular decision. Rulemakers should, however, be prepared for the possibility that using inscrutable machine learning models for some sorts of decision criteria will be normatively unacceptable to the citizenry, regardless of how well-validated the machine learning model might be.

IV. EXPLANATION FOR RULEMAKING

This concluding Part considers how to obtain as many of the traditional benefits of explanation as possible for decision systems that incorporate machine-learning-based decision tools. As noted earlier, machine learning's now-canonical "explainability" problem pertains only to the model's computational mapping between features and outcome variables.[149] While this inscrutability is significant, many societally significant aspects of the development of a machine-learning-based decision tool, its final form, and its integration into a decision system are just as explainable as conventional narrative rules and their underpinnings. In particular, the choice to employ a machine-learning-based decision tool to evaluate particular decision criteria is fully explainable and has significant normative and policy implications that should be open to scrutiny.

Section IV.A thus argues for applying traditional explanation requirements to the explainable aspects of such systems. Section IV.B focuses on the less-discussed issues of communication and coordination within the rulemaking entity, pointing out that these issues require more attention when automated decision tools are introduced because of the

[148] See supra Part II.

[149] See supra section I.B.2.

disciplinary barriers between subject matter experts and data scientists within the rulemaking entity. Section IV.C suggests mechanisms for improving the capacity for rulemaking entities and advocacy groups to make full use of the explanations that would be made available to them under the explanation requirements proposed in section IV.A.[150] While large rulemaking entities and advocacy groups may have sufficient resources to obtain the fairly minimal data science expertise necessary for this purpose, smaller rulemaking entities and advocacy groups might consider pooling resources—though perhaps not with one another—to gain access to it.

A. *Explaining the Incorporation of Machine-Learning-Based Decision Tools*

What sort of explanation should be required when a machine-learning-based decision tool is incorporated into a decision system? In particular, how should this question be answered when the tool is incorporated into decision criteria that operate as a "rule" under the APA?

When a conventional narrative rule is published in the Federal Register for comment, the public receives full notice of its terms.[151] For rule-like criteria, publication allows the public to determine—and critique—how cases of any imaginable sort would be handled.[152] Because inscrutable machine-learning-based decision tools cannot be summarized in narrative form (or even in understandable mathematical or graphical form), there is no way to provide an equivalently detailed mapping from cases to outcomes. If notice and comment demands this sort of detailed mapping, inscrutable decision tools simply cannot be incorporated into APA rules.

While "just say no" to inscrutable decision tools is certainly an appropriate approach in some decision contexts, we should be wary of adopting it as a general response to notice and comment requirements or other explanation mandates. Because machine-learning-based decision tools are attractive to policymakers,[153] an overly expansive interpretation of *what* explanation requires might backfire by motivating rulemakers and courts to adopt narrower interpretations of *whether* such requirements apply at all. Moreover, preemptively depriving society of all such tools for all purposes in all significant decision contexts seems questionable as a policy matter, given the advantages of machine-learning-based decision tools in some contexts.

[150] Proposals for algorithmic impact assessments would produce similar results. See, e.g., Dillon Reisman, Jason Schultz, Kate Crawford & Meredith Whittaker, AI Now Inst., Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability 7–20 (2018), https://ainowinstitute.org/aiareport2018.pdf [https://perma.cc/A3QZ-PCY6]; Andrew D. Selbst, Disparate Impact in Big Data Policing, 52 Ga. L. Rev. 109, 169–82 (2017); Selbst & Barocas, supra note 4, at 1133–38.

[151] Guide to Rulemaking, supra note 72.

[152] See id.

[153] See supra note 18 and accompanying text.

An opposite approach, which may be closer to what is happening on the ground,[154] is to pretend that machine-learning-based decision tools are not really rules at all but something else that does not have to be explained.[155] This approach is, if anything, worse because it deprives society of the benefits of explanation for the aspects that can be explained and ignores the true rule-like nature of the tools.

The analysis here suggests an intermediate approach: Define what constitutes an adequate explanation of a machine-learning-based decision tool and require such an explanation, thus subjecting the incorporation of inscrutable machine learning models to scrutiny while not barring it entirely. This section proposes a framework for adequate explanation composed of two parts: (1) information required to *describe the rule* and (2) information treated as part of the *record* for backing up the rule, as in hard look review.[156] Standard administrative law requirements of notice and recordkeeping could be interpreted in these terms.

1. *Describing the Rule.* — An adequate description of machine-learning-based decision criteria—as would be published in the Federal Register for notice and comment—would include all of the "explainable components" of the rule.[157] Those components are part and parcel of the decisionmaking rule and should be treated as such. They are no more difficult to explain or to understand than conventional narrative rules, and their disclosure fulfills the intended functions of explanation requirements.[158] When disclosure and explanation of a rule is legally re-

[154] The opacity surrounding current government practices makes it difficult to know precisely how these tools are treated. See, e.g., David Curie, AI in the Regulatory State: Stanford Project Maps the Use of Machine Learning and Other AI Technologies in Federal Agencies, Thomson Reuters (June 20, 2019), https://blogs.thomsonreuters.com/answerson/ai-in-the-regulatory-state/ [https://perma.cc/L4QV-UK6G] (noting the relative differences in the integration of artificial intelligence across government agencies); Colin Lecher, New York City's Algorithm Task Force Is Fracturing, The Verge (Apr. 15, 2019), https://www.theverge.com/2019/4/15/18309437/new-york-city-accountability-task-force-law-algorithm-transparency-automation [https://perma.cc/3LNR-U67A] (noting the lack of transparency around the work of the algorithm task force in New York City).

[155] In sentencing proceedings, for example, recidivism risk assessments seem to be treated as a sort of factual evidence. See, e.g., State v. Loomis, 881 N.W.2d 749, 772 (Wis. 2016) (holding that the use of a risk assessment tool at sentencing did not violate defendant's due process rights). For a critique of the treatment of recidivism risk assessment in *Loomis*, see generally Anne L. Washington, How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate, 17 Colo. Tech. L.J. 131 (2018) (arguing for a "new form of reasoning . . . to explain and justify algorithmic results"). For a critique of the treatment of predictive risk assessments as evidence, see generally Steve T. Mckinlay, Evidence, Explanation and Predictive Data Modelling, 30 Phil. & Tech. 461 (2017) ("[T]he claim that [predictive risk models] provide anything close to epistemically justified evidence in a traditional philosophical sense is dubious at best.").

[156] Gillis and Simons come to similar conclusions under the GDPR, though by a different path. See Gillis & Simons, supra note 9 (manuscript at 20–26).

[157] See supra section III.A.1.

[158] See supra section III.A.1.

quired, trade secrecy should not excuse explanation of these aspects, which reflect critical, policy-relevant rulemaking choices.

2. *The Rulemaking Record.* — Selecting training data and validating the tool's performance bear the same relationship to developing a machine-learning-based decision tool that more familiar sorts of factual inquiry and statistical analysis bear to the development and justification of a conventional rule. Summary information about the training data, explanations of how it was sourced, descriptions of the validation process, and validation results should thus be part of the rulemaking record and made available on the same terms as other parts of the record. Rulemaking entities should not sign confidentiality agreements regarding this information. However, the training data set itself should ordinarily be kept confidential for privacy reasons. Confidentiality agreements regarding certain technical parameters and details of the machine learning process might also be appropriate.

B. *Coordination and Communication Between Data Scientists and Substantive Rulemakers*

To promote informed, effective overall decision-criteria design, data scientists need a deep understanding of both the overall goals of the decision system and how the criteria evaluated by the automated tools will be incorporated into the ultimate decision. Concomitantly, substantive rulemakers need to acquire a basic understanding of the machine learning process so that they can make appropriate choices about whether to automate particular decision criteria, interact meaningfully with data scientists throughout the development process, and design appropriate combination schemes for adjudicators to follow when using the outputs of machine-learning-based tools.

The incentives provided by the proposed explanation requirements will go some way toward facilitating the requisite communication and coordination between data scientists and substantive experts. But because these interactions are of such vital importance to decision quality and face significant barriers, more may be necessary. Rulemakers who are considering incorporating a machine-learning-based tool into a decision system would therefore be well advised to adopt a prospective "by design" plan aimed at ensuring the necessary level of cooperation and communication between data scientists and substantive rulemakers.

Ideally, the development of machine-learning-based decision tools would be brought in-house, so that dedicated data scientists could develop substantive expertise to support their work. That approach is probably overly ambitious for most rulemaking entities, who would only be undertaking such development on a sporadic basis. Larger entities should still consider hiring an in-house data scientist whose role would be not only to advise the rulemaking entity in its interactions with outside contractors but also to initiate and facilitate the necessary close interactions between rulemakers and outside data scientists.

At a minimum, where an automated decision tool is procured from outside data scientists, substantive rulemakers must demand clear and thorough explanations of the aspects described in section III.A so that they can understand the outputs of the automated decision tool and create appropriate combination schemes for adjudicators to use. The depth of information that is available about the automated tool constrains the sorts of combination schemes that adjudicators can implement. Clear communication and coordination between data scientists and substantive rulemakers are critical to assessing the severity of those constraints. As discussed above, inscrutability is especially likely to limit the extent to which adjudicators can serve the role in addressing generalizability issues that is commonly assigned to them in conventional decision systems. Rulemakers must understand and confront these and other trade-offs involved in using inscrutable decision tools.

C. *Centers of Data Science Expertise for Rulemaking Entities and Advocacy Groups*

While larger rulemaking entities may be able to develop data science expertise to help them communicate and coordinate with the data science contractors who will probably continue to do most development of machine-learning-based decision tools, smaller rulemaking entities will likely be strapped to find the necessary resources. This is a problem because smaller entities are perhaps most likely to be attracted to the potential cost-savings of automation, while also being least able to afford to acquire data science expertise, increasing the temptation to use off-the-shelf solutions. It would be wise for smaller rulemaking entities to develop mechanisms for pooling resources with similarly situated entities to provide access to data science expertise. Ideally, such pooling would bring rulemaking entities in similar substantive arenas together so that the data scientists they work with could also build up substantive expertise. This proposal is tentative; its feasibility would depend on working out the details. If this sort of resource pooling is feasible, it should receive public support, and participation might even be mandated. Advocacy groups and other agenda setters in a given arena may also benefit from creating similar centers of data science expertise to assist them in understanding the explanations provided by rulemakers and ensuring accountability.

CONCLUSION

Delegated, distributed decision systems—which are responsible for many highly consequential decisions affecting individuals—confront issues of cost, efficiency, and consistency that make automated decision tools particularly attractive. Though scholars and policymakers have focused on explanations to decision subjects and accountability to the public, the inscrutability of automated decision tools has significant, and underappreciated, implications for the explanatory flows required to develop and implement such systems. While sharing the explainable aspects of these tools can replicate some of explanation's traditional functions, using inscrutable automated decision tools inevitably degrades decision-criteria development in some respects. Thus, in weighing the advantages and disadvantages of such tools for a given decision context, policymakers and system designers should consider how inscrutability affects rulemakers and, as I discuss elsewhere, adjudicators, along with its direct impact on decision subjects.

MINDS, MACHINES, AND THE LAW: THE CASE OF VOLITION IN COPYRIGHT LAW

Mala Chatterjee & Jeanne C. Fromer***

The increasing prevalence of ever-sophisticated technology permits machines to stand in for or augment humans in a growing number of contexts. The questions of whether, when, and how the so-called actions of machines can and should result in legal liability thus will also become more practically pressing. One important set of questions that the law will inevitably need to confront is whether machines can have mental states, or—at least—something sufficiently like mental states for the purposes of the law. This is because a number of areas of law have explicit or implicit mental state requirements for the incurrence of legal liability. Thus, in these contexts, whether machines can incur legal liability turns on whether a machine can operate with the requisite mental state. Consider the example of copyright law. Given the long history of mechanical copying, courts have already faced the question of whether a machine making a copy can have the mental states required for liability. They have often answered with a resounding, unconditional "no." But this Essay seeks to challenge any generalization that machines cannot operate with a mental state in the eyes of the law. Taking lessons from philosophical thinking about minds and machines—in particular, the conceptual distinction between "conscious" and "functional" properties of the mind—this Essay uses copyright's volitional act requirement as a case study to demonstrate that certain legal mental state requirements might seek to track only the functional properties of the states in question, even ones which can be possessed by machines. This Essay concludes by considering how to move toward a more general framework for evaluating the question of machine mental states for legal purposes.

* Law Clerk for the Honorable Judge Robert D. Sack, United States Court of Appeals for the Second Circuit; Ph.D. Candidate in Philosophy, New York University; Fellow, Engelberg Center for Innovation Law and Policy; Visiting Fellow, Information Society Project, Yale Law School; J.D., New York University School of Law 2018.

** Professor of Law, New York University School of Law. We thank Arnaud Ajdler, Jack Balkin, Barton Beebe, Colin Bradley, Christopher Buccafusco, Dan Burk, David Chalmers, Colleen Chien, Rebecca Crootof, Brian Frye, Eric Goldman, Michael Grynberg, Nathan Gusdorf, Scott Hemphill, Ben Holguín, Bert Huang, Arden Koehler, Andrew Lee, S. Matthew Liao, Jeffrey Lipshaw, Michael Meurer, Liam Murphy, Betsy Rosenblatt, Olga Russakovsky, Matthew Sag, Erick Sam, Jason Schultz, Jule Sigall, Christopher Sprigman, Jeff Stein, Katherine Strandburg, Olivier Sylvain, Jacob Victor, Patrick Winston, Felix Wu, Gideon Yaffe, and participants at workshops at New York University School of Law, Yale Law School, the 2019 Works in Progress in Intellectual Property Colloquium, and the *Columbia Law Review* Symposium on "Common Law for the Age of AI," for choosing to share knowledgeable comments. Thanks to Moses Dyckman for helpful research assistance. Jeanne Fromer gratefully acknowledges support from the Filomen D'Agostino and Max E. Greenberg Research Fund.

INTRODUCTION

With the increasing prevalence of ever more sophisticated technology—which permits machines to stand in for or augment humans in a growing number of contexts—the questions of whether, when, and how the so-called actions of machines can and should result in legal liability will become more practically pressing.[1] Although the law has yet to fully grapple with questions such as whether machines are (or can be) sufficiently humanlike to be the subjects of law, philosophers have long contemplated the nature of machines.[2] Philosophers have considered, for instance, whether human cognition is fundamentally computation—such that it is in principle possible for future artificial intelligences (AI) to possess the properties of human minds, including consciousness, semantic understanding, intention, and even moral responsibility—or if humans and machines are instead fundamentally different, no matter how sophisticated AI becomes.[3] It is thus unsurprising that, in thinking through how the law should accommodate and govern an increasingly AI-filled world, the lessons and frameworks to be gleaned from these philosophical discussions will have undeniable relevance.

One important set of questions that the law will inevitably need to confront is whether machines can have mental states, or—at least—something sufficiently like mental states for the purposes of the law. This is because a wide range of areas of law have explicit or implicit mental state requirements for the incurrence of legal liability.[4] Consider, for example, questions of intent and recklessness versus negligence in tort law; mens rea and actus reus in criminal law; offer and acceptance in contract law; and, as we will see, infringement and authorship in copyright law. In each of these contexts, the law either implicitly or explicitly asks for the presence of some particular mental state on the part of the actors in question. Whether the operations of machines can incur legal liability—and what kind of liability they can incur—would thus often seem to turn on whether a machine is regarded as operating with the mental state required.

In some contexts, the decision already seems to have been made that machines can never possess the mental states required for liability. Consider copyright law's volitional act requirement for infringement. Copyright law has generally claimed that machines making copies of protected material lack the requisite volition for this conduct to give rise to legal liability on the part of those responsible for the machine, even when the machine has been designed to make copies, often of copyrighted works.[5] In other contexts, such as criminal and tort law, the

[1] See infra section I.B.

[2] See infra Part III.

[3] See infra Part III.

[4] See infra section I.A.

[5] See infra Part II.

question of machines' capacity for mental states remains open and underexplored.[6]

This Essay aims to challenge any hasty and blanket generalization that machines cannot have mental states as a legal matter, drawing on philosophical thinking surrounding mental states and using copyright's volitional act requirement as a case study. In so doing, this Essay concludes that—as a matter of copyright doctrine—a copying technology might be sufficiently "volitional" for the technology provider to be held directly liable for the technology's so-called actions in producing copies; and—as a matter of general legal theory—machines in some contexts might be capable of being sufficiently "mental" to count as agents of the humans behind them, depending on the aims of the area of law in question.[7] This conclusion is thus not merely of philosophical interest but one with practical implications for determinations of legal liability. In the context of copyright law, this Essay's chosen case study, this conclusion has implications for who is and is not directly accountable for the copying of protected material and for the law's ability to effectuate its goals of encouraging the creation and dissemination of expressive works.

To mount this Essay's challenge, after giving an overview of mental states in the law and the puzzle raised by technological advancement in Part I, as well as the specific challenges posed by copyright law in Part II, Part III of the Essay recounts two of the most influential philosophical discussions on minds and machines, and the resulting theoretical distinction between the conscious and functional properties of mental states. Using this distinction as a framework, this Essay argues that it is an open question whether the law's mental state requirements seek to track the conscious or merely functional properties of the particular mental state in question,[8] and the analysis depends on the ultimate aims of the relevant area of law. Part IV then defends the view that copyright law's volitional act requirement might be interested in merely functional

[6] See generally Mark A. Geistfeld, A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation, 105 Calif. L. Rev. 1611 (2017) (tort liability); Gabriel Hallevy, "I, Robot—I, Criminal"—When Science Fiction Becomes Reality: Legal Liability of AI Robots Committing Criminal Offenses, 22 Syracuse Sci. & Tech. L. Rep. 1 (2010) (criminal law); Ignatius Michael Ingles, Note, Regulating Religious Robots: Free Exercise and RFRA in the Time of Superintelligent Artificial Intelligence, 105 Geo. L.J. 507, 516 n.67 (2017) (criminal law).

[7] See infra Parts IV–V.

[8] Note that there are arguably nonconscious mental states aside from functional mental states, such as intentional and computational states. In this way, the distinction on which we focus—consciousness versus functionality—is not exhaustive, as one could similarly ask whether the law cares about intentionality, computation, and so forth. Nonetheless, the Essay focuses on consciousness versus functionality not only for the sake of simplicity, but also because this distinction is plausibly the most important one for legal purposes. The Essay otherwise leaves the question of whether intentionality (or other nonconscious, nonfunctional properties of mental states) should ever matter to the law for exploration in future work.

properties, which could—in principle—be replicated by machines. Next, Part V considers which functional properties copyright law might seek to track and what a machine might have to look like to be "functionally volitional" under copyright law, to count as the technology provider's agent, and thereby to give rise to direct liability. These relevant functional properties include the ability to pause and analyze the nature of the work in question before "choosing" to undertake an act of copying, one which might cause exposure to liability. On the basis of this framework, this Essay concludes that machines with the appropriate functionality might satisfy copyright law's volitional act requirement, thus forming the basis for holding technology providers directly liable for infringement. Finally, generalizing this Essay's framework, Part VI offers preliminary thoughts on machines and mental state requirements in the contrasting contexts of criminal law and copyright authorship doctrine, as well as a general hypothesis regarding when the law is interested in conscious versus merely functional properties of the mental states in question.

I. MENTAL STATES, TECHNOLOGY, AND THE LAW

This Part explores the intersection of mental states and technology under the law. It first provides an overview of the law's mental state requirements, and then surveys how businesses might use machines in lieu of humans to perform various operations that could—or would—incur liability if performed by a human, such that technological advancement inevitably raises the legal question of machine mental states.

A. *Mental State Requirements in the Law*

Mental state requirements for legal liability are pervasive. The most familiar include requirements of purpose (or intent), knowledge, and recklessness, in contrast to negligence (which is not itself a mental state but might be understood as distinguishable from, say, recklessness by the absence of such a state).[9] Each relates in differing ways to beliefs or desires.[10] Volition—which might be defined as the cause of willful actions, and which thus distinguishes actions from involuntary bodily

[9] See, e.g., Model Penal Code § 2.02 (Am. Law Inst. 1985); see also Kyron Huigens, On Commonplace Punishment Theory, 2005 U. Chi. Legal F. 437, 453 ("In negligence and the other non-intentional fault doctrines, fault is found not in a discrete mental state, but in a broader set of facts surrounding the offense.").

[10] See Kenneth W. Simons, Rethinking Mental States, 72 B.U. L. Rev. 463, 464–65 (1992) ("Properly understood, the principal mental state concepts do not reflect a single hierarchy of legal significance. Rather, they conceal two distinct mental state hierarchies, of desire and belief, as well as a third hierarchy, of conduct, which does not essentially involve mental states.").

movements[11]—can be understood as a mental state as well.[12] Thus, in addition to any further mens rea requirements, any area of law requiring a willful action for liability is implicitly asking for a mental state as well, because the presence of volition is what makes a movement count as a willful action (rather than, say, a muscle spasm) in the first place.[13] Mental state requirements thus exist in nearly every area of law, including criminal law, torts, and contract.[14] Indeed, these requirements are so prevalent that there is even a legal category arguably defined in terms of an absence of any mens rea beyond volition itself: namely, strict liability.[15] These requirements are premised on the assumption that the mind—and not just the body—matters to the law.[16] In other words, when such requirements exist, the body might move to do something prohibited, but only when this is conjoined with the corresponding illicit mental state is this a prohibited action.

As an evidentiary matter, discerning the presence of a mental state in a human requires "mind reading," so to speak, because people cannot directly observe or measure a mental state.[17] Nonetheless, the law typically feels comfortable—though perhaps it should not[18]—answering the question of whether a human had the required mental state. In light of these requirements, as machines become more pervasive in performing operations that humans traditionally performed, the law will find itself needing to assess not just the permissibility of machines' operations but also whether they have operated with an illicit mental state.

[11] See Michael S. Moore, Act and Crime: The Philosophy of Action and Its Implications for Criminal Law 113–65 (1993) [hereinafter Moore, Act and Crime] (defending a theory of volition as the mental state that causes actions).

[12] See, e.g., id. at 115 ("'Volition' names a state or an event within the mind of the actor.").

[13] See id. at 113–65.

[14] See, e.g., Kent Greenawalt, A Pluralist Approach to Interpretation: Wills and Contracts, 42 San Diego L. Rev. 533, 575–82 (2005) (contracts); Simons, supra note 10, at 468–73 (criminal law and torts).

[15] See Simons, supra note 10, at 464.

[16] See, e.g., Keren Shapira-Ettinger, The Conundrum of Mental States: Substantive Rules and Evidence Combined, 28 Cardozo L. Rev. 2577, 2579–81 (2007) ("[C]riminal law has adopted the vague metaphysical dualistic vision between a forbidden act and a state of mind that accompanied it."). Some have criticized this assumption, suggesting it ought to be replaced with an integrated actus reus and mens rea. See, e.g., Douglas N. Husak, Philosophy of Criminal Law 126 (1987) (advocating for this integration "as an indivisible product of both what one thinks and what one does").

[17] See Teneille Brown & Emily Murphy, Through a Scanner Darkly: Functional Neuroimaging as Evidence of a Criminal Defendant's Past Mental States, 62 Stan. L. Rev. 1119, 1129–30 (2010) ("Because we cannot presently read someone's mind to determine her mens rea at the time of the crime, the jury is often told it can rely on the objective circumstances surrounding the criminal's conduct to draw inferences about her state of mind.").

[18] See, e.g., James A. Macleod, Belief States in Criminal Law, 68 Okla. L. Rev. 497, 502–03, 514–34 (2016) (drawing on experimental epistemology to criticize how juries likely decide on the presence of a mental state).

B. *The Present and Future of Technology*

Increasingly, tasks once performed by only humans are being carried out or augmented by machines, which often perform better than humans ever could. In the copyright space alone—on which the Essay elaborates in the next Part—there are devices that can now recognize songs and other expressive content by listening to them,[19] virtual assistants and bots that can locate and play user-requested content,[20] and software that can use machine learning techniques to create artwork based on a model derived from 15,000 portraits painted over the past six centuries.[21] A piece of art created using this software recently sold at auction for over $400,000.[22]

Thus, questions of so-called machine liability are becoming more pressing. Legal scholars have already been puzzling over a tort liability regime for self-driving cars.[23] Plausibly, we might soon find ourselves asking whether a bot producing defamatory content about a public figure can itself have actual malice; whether an algorithm assessing risk can have discriminatory intent; or whether the price-setting systems of competing businesses can collude from the perspective of antitrust law. And in the copyright space, we might wonder whether technology creators or owners can be directly liable for copyright infringement when a bot fetches an infringing copy of a song in response to a user's request for that song or when software taught on portraits produces an artwork that is copied from and substantially similar to an existing portrait on which the software was trained.

[19] E.g., Trent Gillies, Shazam Names That Tune, Drawing in Money and Users, CNBC (June 14, 2015), https://www.cnbc.com/2015/06/14/shazam-names-that-tune-drawing-in-money-and-users.html [https://perma.cc/6DEK-L2KV].

[20] E.g., Taylor Martin, 9 Alexa Tips for Music Lovers, CNET (Jan. 22, 2019), https://www.cnet.com/how-to/alexa-tips-for-music-lovers [https://perma.cc/4ATW-C6HX].

[21] Is Artificial Intelligence Set to Become Art's Next Medium?, Christie's (Dec. 12, 2018), https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx [https://perma.cc/R6NB-RU5F].

[22] Id.

[23] See, e.g., Geistfeld, supra note 6, at 1691–94 (arguing that a combination of state products liability law and federal regulations can provide an effective framework for self-driving cars); Gary E. Marchant & Rachel A. Lindor, The Coming Collision Between Autonomous Vehicles and the Liability System, 52 Santa Clara L. Rev. 1321, 1335–39 (2012) (suggesting "legal and policy tools that may help protect manufacturers [of autonomous vehicles] from liability," including the assumption of risk defense, legislative limitations on liability, and federal preemption of state tort actions); Bryant Walker Smith, Automated Driving and Product Liability, 2017 Mich. St. L. Rev. 1, 2 ("[T]he current product liability regime, while imperfect, is probably compatible with the adoption of automated driving systems."); Harry Surden & Mary-Anne Williams, Technological Opacity, Predictability, and Self-Driving Cars, 38 Cardozo L. Rev. 121, 178–80 (2016) (describing the potential for tort liability to encourage autonomous car manufacturers to program more predictable movements, as well as the ability for autonomous cars to transform the issue of fault in car accidents by providing a "'black box' record").

II. THE COPYRIGHT EXAMPLE: A LONG HISTORY OF MECHANICAL COPYING

This Part uses copyright infringement as this Essay's case study for the challenge posed for the law by mental states and machines. In particular, this Part recounts copyright law's extensive history of mechanical copying, which has long provoked courts to explore whether and when machines and their owners can be directly liable for infringement. This history has led courts to develop a volitional act requirement for copyright infringement, while suggesting that this requirement—though *always* satisfied by human actions—can never be satisfied by machines. This Part also explains why the volitional act requirement ought to be understood as a mental state. For these reasons, the requirement provides a good test bed to explore whether machines should ever possess mental states as a legal matter.

A. *Background*

By way of background, American copyright law protects "original works of authorship fixed in any tangible medium of expression," including literary works, sound recordings, and movies.[24] A copyright holder receives, among other things, the exclusive right to reproduce the work, distribute copies of it, and prepare derivative works,[25] typically until seventy years after the author's death.[26] Copyright protection extends to the expression of particular ideas rather than to the ideas themselves.[27] Yet protection actually reaches well beyond the literal work to works that are copied and substantially similar,[28] "else a plagiarist would escape by immaterial variations."[29]

The most widely embraced theory of copyright law in America is utilitarian and, in particular, economic.[30] According to this theory,

[24] 17 U.S.C. § 102(a) (2012).

[25] Id. § 106.

[26] Id. § 302(a).

[27] See id. § 102(b); Nichols v. Universal Pictures Corp., 45 F.2d 119, 121 (2d Cir. 1930).

[28] Corwin v. Walt Disney Co., 475 F.3d 1239, 1253 (11th Cir. 2007) (citing Herzog v. Castle Rock Entm't, 193 F.3d 1241, 1249 (11th Cir. 1999)).

[29] *Nichols*, 45 F.2d at 121.

[30] See, e.g., Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 558 (1985) (embracing an economic theory of copyright, and stating that "[b]y establishing a marketable right to the use of one's expression, copyright supplies the economic incentive to create and disseminate ideas"); Shyamkrishna Balganesh, Foreseeability and Copyright Incentives, 122 Harv. L. Rev. 1569, 1576–77 (2009) ("[C]opyright law in the United States has undeniably come to be understood almost entirely in utilitarian, incentive-driven terms."); Jeanne C. Fromer, Expressive Incentives in Intellectual Property, 98 Va. L. Rev. 1745, 1750–52 (2012) ("The Supreme Court, Congress, and many legal scholars consider utilitarianism the dominant purpose of American copyright and patent law."); William M. Landes & Richard A. Posner, An Economic Analysis of Copyright Law, 18 J. Legal Stud. 325, 326 (1989) (proposing an "economic model of copyright protection").

copyright law provides the incentive of exclusive rights for a limited duration to authors to motivate them to create and distribute culturally valuable works.[31] Without this incentive, the theory goes, authors might not invest the time, energy, and money necessary to create and distribute these works because they might be copied cheaply and easily by free riders, eliminating authors' ability to profit from their works.[32] By allowing a copyright holder to recover damages from and enjoin an infringer that breaches the copyright holder's exclusive rights—thereby undermining copyright's pecuniary incentive—the law preserves the copyright incentive.[33]

A utilitarian theory of copyright law rests on the premise that the benefit to society of creators crafting valuable works offsets the costs to society of the incentives the law offers to creators.[34] To prevent excessive rights that would undercut the goals of dissemination of works and of creation that builds on preexisting works, copyright law therefore limits copyright's duration and scope in certain ways.[35] For example, copyright law excuses some third-party uses that would otherwise be infringing by deeming them to be "fair use."[36] The fair use doctrine enables third parties to create culturally valuable works that must borrow from the original work in some capacity in order to succeed, often transforming it.[37]

Moreover, copyright infringement is understood to be a strict liability offense. At the extreme, a person can infringe another's copyright even if they copy from the third party's work without any awareness of the

[31] Stewart E. Sterk, Rhetoric and Reality in Copyright Law, 94 Mich. L. Rev. 1197, 1197 (1996).

[32] See id.

[33] See Roger D. Blair & Thomas F. Cotter, An Economic Analysis of Damages Rules in Intellectual Property Law, 39 Wm. & Mary L. Rev. 1585, 1617–46 (1998) ("[A] simple model of intellectual property rights suggests that the prevailing plaintiff in a . . . copyright . . . infringement action should be able to recover the greater of her lost profit attributable to the infringement, or the defendant's profit so attributable . . . ."); Jeanne C. Fromer & Mark A. Lemley, The Audience in Intellectual Property Infringement, 112 Mich. L. Rev. 1251, 1299–1300 (2014) (discussing the "multiple vantage points" used when assessing a copyright infringement as a way to structure when there is infringement liability and thus preserve copyright's incentive).

[34] See Mark A. Lemley, The Economics of Improvement in Intellectual Property Law, 75 Tex. L. Rev. 989, 996–97 (1997).

[35] See id. at 996–98.

[36] 17 U.S.C. § 107 (2012).

[37] See Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 577 (1994) ("The fair use doctrine thus 'permits [and requires] courts to avoid rigid application of the copyright statute when, on occasion, it would stifle the very creativity which that law is designed to foster.'" (alteration in original) (quoting Stewart v. Abend, 495 U.S. 207, 236 (1990))); Pierre N. Leval, Toward a Fair Use Standard, 103 Harv. L. Rev. 1105, 1111–16 (1990) ("Quotation can be vital to the fulfillment of the public-enriching goals of copyright law. The first fair use factor calls for a careful evaluation whether the particular quotation is of the transformative type that advances knowledge and the progress of the arts . . . .").

fact that they have copied.[38] For example, singer Michael Bolton was found liable for infringement for subconsciously copying the Isley Brothers' song "Love Is a Wonderful Song" decades later in his song of the same name.[39] As Judge Learned Hand explained,

> Everything registers somewhere in our memories, and no one can tell what may evoke it. . . .
>
> . . . Once it appears that another has in fact used the copyright as the source of his production, he has invaded the author's rights. It is no excuse that in so doing his memory has played him a trick.[40]

B. *The Player Piano Roll*

In light of a consistent stream of advancements in copying technologies, copyright law has already had to grapple with whether and when copies made by machines constitute copyright infringement.[41] One of the most striking illustrations of this dates back to the early twentieth century, when copyright law faced player piano rolls: rolls of paper with perforations in accordance with musical works.[42] When installed on a player piano, these rolls cause the piano to play notes in sequence as determined by the position and length of the perforations, thereby performing the song encoded therein. In 1908, the Supreme Court considered in *White-Smith Music Publishing Co. v. Apollo Co.* whether the piano rolls—which would be "read" by a machine to play the encoded musical composition rather than by a human—were "copies" of the musical composition, thereby constituting copyright infringement.[43] The plaintiff in the case owned copyrights in certain musical compositions, and the

[38] See, e.g., Three Boys Music Corp. v. Bolton, 212 F.3d 477, 482–85 (9th Cir. 2000) (stating that "[s]ubconscious copying has been accepted" alongside proof of widespread dissemination to satisfy proof of the reasonable access element of copyright infringement); ABKCO Music, Inc. v. Harrisongs Music, Ltd., 722 F.2d 988, 998–99 (2d Cir. 1983) ("It is not new law in this circuit that when a defendant's work is copied from the plaintiff's, but the defendant in good faith has forgotten that the plaintiff's work was the source of his own, such 'innocent copying' can nevertheless constitute an infringement.").

[39] *Three Boys Music*, 212 F.3d at 484–85.

[40] Fred Fisher, Inc. v. Dillingham, 298 F. 145, 147–48 (S.D.N.Y. 1924).

[41] Copyright law would likely not exist in the first place without the printing press, which made the large-scale copying of written material plausible. See Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 430 (1984) ("Indeed, it was the invention of a new form of copying equipment—the printing press—that gave rise to the original need for copyright protection.").

[42] Zhengshan Shi, Kumaran Arul & Julius O. Smith, Modeling and Digitizing Reproducing Piano Rolls, *in* Proceedings of the 18th International Society for Music Information Retrieval Conference 197, 197 (Xiao Hu, Sally Jo Cunningham, Doug Turnbull & Zhiyao Duan eds., 2017), https://ismir2017.smcnus.org/wp-content/uploads/2017/10/25_Paper.pdf [https://perma.cc/W9EH-JLJL].

[43] 209 U.S. 1, 17–18 (1908). Under the copyright statute in place at the time—and continuing through its current version—copyright law deemed copying of copyrighted works to be infringement. Id. at 9.

defendant was in the business of making and selling player pianos and piano rolls.[44]

The Supreme Court ultimately held that the piano roll was not a copy of the musical composition it represented (and therefore the plaintiff could not prohibit this type of reproduction by the defendant).[45] In particular, the Court reasoned that something could not count as an infringing use unless it was "put in a form which [humans] can see and read."[46] Because people did not read piano rolls as they read sheet music, piano rolls did not satisfy this requirement. The Court thought it irrelevant that "[t]hese perforated rolls are parts of a machine which, when duly applied and properly operated in connection with the mechanism to which they are adapted, produce musical tones in harmonious combination."[47]

In its ruling, the Court thus adopted the view that machines were unlike humans for purposes of copyright infringement: Machine-read materials did not constitute copyright infringement unless humans can read the same material as well.[48] However, Congress evidently did not share the Supreme Court's broad view on this distinction between humans and machines.[49] Although there are arguably justifications for a focus on human readability, *White-Smith*'s formalism provoked severe criticism.[50] Even if a person could not read or hear the musical composition encoded in a piano roll, that same person could still consume the work with the help of a player piano.[51] As a practical matter, *White-Smith* meant that copiers could circumvent copyright protections by creating copies of a work that were unreadable by humans, but could be made comprehensible with the aid of a machine.[52]

The following year, Congress overturned the specific holding of *White-Smith* by granting copyright holders in musical works the right to control the mechanical reproduction of their works and instituting a compulsory license scheme for manufacturers of piano rolls and other

[44] Id. at 8–9.

[45] Id. at 18.

[46] Id. at 17.

[47] Id. at 18.

[48] Id. at 17–18.

[49] See Yvette Joy Liebesman, Redefining the Intended Copyright Infringer, 50 Akron L. Rev. 765, 790 (2016) (stating that "Congress amended the Copyright Act to include these works under its purview" (citing An Act to Amend and Consolidate the Acts Respecting Copyright, ch. 320, 35 Stat. 1075, 1081–82 (1909)).

[50] See, e.g., H.R. Rep. No. 94-1476, at 52 (1976) (criticizing *White-Smith* for its "artificial and largely unjustifiable distinction[] . . . under which statutory copyrightability . . . has been made to depend upon the form or medium in which the work is fixed").

[51] *White-Smith*, 209 U.S. at 8–10.

[52] See Liebesman, supra note 49, at 787–90 (finding that the Supreme Court's decisions "confin[ing] copies of musical works . . . to those specific mediums of expression defined by Congress . . . resulted in a larger reach of legal copying and subsequently a smaller cohort of who was an intended infringer").

mechanical reproductions.[53] And almost seventy years later, Congress changed its definition for copyright law of "copies" to include not only "material objects" that can be read or perceived "directly" by humans but also those "from which the work can be perceived, reproduced, or otherwise communicated . . . with the aid of a machine or device."[54] With that definition, Congress took an expansive view of machine-readable forms of works as "copies," so long as humans could perceive or read them via the machine.

C. *The Internet*

Nonetheless, further questions as to machines' ability to engage in copyright infringement subsequently arose, especially as the internet era dawned in the 1990s. For the first time, machines—computers—interconnected on a vast network around the world were copying and transmitting material to one another (and ultimately often to people using these machines). Any human posting or emailing material that infringed another's copyright would therein provoke countless interconnected machines to make copies of this material as well. Some frustrated copyright holders sued certain of these users and machine owners—typically, internet service providers—for copyright infringement.

The foundational case of *Religious Technology Center v. Netcom On-Line Communication Services, Inc.* addressed the liability of internet server owners.[55] *Netcom* was a suit by the Church of Scientology against both former minister Dennis Erlich, for uploading messages to Usenet containing copyrighted church texts and criticism of the church, and internet service providers, including BBS and Netcom, whose servers created copies of those messages.[56] The Northern District of California viewed the liability of the entities deploying these servers as turning on "whether possessors of computers are liable for incidental copies automatically made on their computers using their software as part of a process initiated by a third party."[57] But the court refused to assign liability to the server owners: "Although copyright is a strict liability statute, there should still be some element of volition . . . which is lacking where a defendant's system is merely used to create a copy by a third party."[58] The *Netcom* court thought that because the defendants' "systems can operate without any human intervention, . . . the mere fact that Netcom's system incidentally makes temporary copies of [the church's] works does not

[53] An Act to Amend and Consolidate the Acts Respecting Copyright § 1(e).

[54] Copyright Act of 1976, Pub. L. No. 94-553, § 101, 90 Stat. 2541, 2542 (codified as amended at 17 U.S.C. § 101 (2012)).

[55] 907 F. Supp. 1361 (N.D. Cal. 1995).

[56] See id. at 1365–66.

[57] Id. at 1368.

[58] Id. at 1370.

mean Netcom has caused the copying."[59] The court emphasized the risk of establishing a contrary rule:

> [A contrary rule] would also result in liability for every single [internet] server in the worldwide link of computers transmitting [the ex-church minister's] message to every other computer. These parties, who are liable under [the church's] theory, do no more than operate or implement a system that is essential if [internet] messages are to be widely distributed. There is no need to construe [copyright law] to make all of these parties infringers.[60]

Thus, the *Netcom* court strongly suggested that—although a human using a machine to make a copy is thereby volitionally infringing a copyright—a machine itself cannot possess the requisite volition to be regarded as an infringer, or as thereby "acting" on behalf of the technology provider.[61]

Building on *Netcom* and its progeny,[62] the Second Circuit further stressed the differential treatment of humans and machines with regard to volition and copyright infringement in *Cartoon Network LP v. CSC Holdings, Inc.*[63] In that case, the court held that a cable company's remote-storage digital video recording system did not directly infringe the copyrights of a cable television company when cable company customers requested or played back recordings on this system.[64] For one thing, the court dismissed the possibility that the cable company satisfied the volitional act requirement for infringement liability by virtue of its "conduct in designing, housing, and maintaining a system that exists only to produce a copy . . . made automatically upon [a] customer's command."[65] For even though the copying was instrumental to the function of the recording system, the court held that it was the customer requesting the recording—rather than the system or its owner—who made the copy.[66] The court thought that it would have been a different situation, however, had the customer requested a human employee of the cable system—rather than the machine itself—to make the copy: "In determining who actually 'makes' a copy, a significant difference exists between making a request to a human employee, who then volitionally operates the copying system to make the copy, and issuing a command

[59] Id. at 1368–69.

[60] Id. at 1369–70. The court left open the possibility that the internet service providers would instead be liable for contributory infringement. Id. at 1369, 1373–75.

[61] See id. at 1370.

[62] Cases in the intervening years on this issue include CoStar Grp., Inc. v. LoopNet, Inc., 373 F.3d 544 (4th Cir. 2004); Field v. Google Inc., 412 F. Supp. 2d 1106 (D. Nev. 2006); Playboy Enters., Inc. v. Russ Hardenburgh, Inc., 982 F. Supp. 503 (N.D. Ohio 1997); Marobie-FL, Inc. v. Nat'l Ass'n of Fire Equip. Distribs., 983 F. Supp. 1167 (N.D. Ill. 1997).

[63] 536 F.3d 121 (2d Cir. 2008).

[64] See id. at 123.

[65] Id. at 131.

[66] Id.

directly to a system, which automatically obeys commands and engages in no volitional conduct."[67] The Second Circuit seemed to state categorically that machines—circa 2008—always lack the requisite volition to be infringers acting on behalf of technology providers, whereas humans, including human employees, always possess it.[68]

While some courts were denying the possibility that machines could volitionally infringe on behalf of technology providers, others seemed to ignore the volitional act requirement entirely, instead readily assuming—without analysis—that computers' owners had infringed when their machines automatically copied protected content. For example, in a series of cases, courts generally found businesses operating search engines not liable for copying infringing works found online to index and make them available for user searching.[69] But these courts never paused to question whether the machines had *volitionally* copied, proceeding instead to decide that there was in fact a prima facie case of copyright infringement by the search engine operators but that their copying was nonetheless fair use.[70] Similarly, the Supreme Court, in *American Broadcasting Companies v. Aereo, Inc.*, made no mention of volition before finding the owner of many small internet-connected antennae liable for streaming (that is, publicly performing) broadcast television programming to subscribers.[71]

[67] Id.

[68] In some ways, a prior decision by the Fourth Circuit had already muddied the volition waters further. The Fourth Circuit found that an internet service provider lacked volition when the company had its human employees take a quick look at whether commercial real estate photographs posted by users seemed to infringe on third parties' copyrighted material and its computers copied the infringing material to check it against any new material uploaded by that user. See CoStar Grp. v. LoopNet, Inc., 373 F.3d 544, 556 (4th Cir. 2004). The court elaborated:

> The employee's look is so cursory as to be insignificant, and if it has any significance, it tends only to lessen the possibility that [the provider]'s automatic electronic responses will inadvertently enable others to trespass on a copyright owner's rights. In performing this gatekeeping function, [the provider] does not attempt to search out or select photographs for duplication; it merely *prevents* users from duplicating certain photographs. . . . [The provider] can be compared to an owner of a copy machine who has stationed a guard by the door to turn away customers who are attempting to duplicate clearly copyrighted works. [The provider] has not by this screening process become engaged as a "copier" of copyrighted works who can be held liable under . . . the Copyright Act.

Id.

[69] See Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1176–77 (9th Cir. 2007); Kelly v. Arriba Soft Corp., 336 F.3d 811, 822 (9th Cir. 2003).

[70] *Perfect 10*, 508 F.3d at 1168; *Kelly*, 336 F.3d at 822. Perhaps the courts never considered volition because the machines' owners in these cases provoked the copying in the first instance. Cf. Robert C. Denicola, Volition and Copyright Infringement, 37 Cardozo L. Rev. 1259, 1279–80 (2016) ("[I]f no third party has participated in the alleged infringement, defendants rarely invoke the volition requirement; when they do, the issue is quickly resolved in favor of the plaintiffs.").

[71] See 134 S. Ct. 2498, 2498–511 (2014).

In dissent, Justice Scalia lambasted the majority for failing to consider whether volition was present as a prerequisite to finding infringement:

> Although we have not opined on the issue, our cases are fully consistent with a volitional-conduct requirement. . . .
>
> The volitional-conduct requirement is not at issue in most direct-infringement cases; the usual point of dispute is whether the defendant's conduct is infringing (*e.g.*, Does the defendant's design copy the plaintiff's?), rather than whether the defendant has acted at all (*e.g.*, Did this defendant create the infringing design?). But it comes right to the fore when a direct-infringement claim is lodged against a defendant who does nothing more than operate an automated, user-controlled system. Internet-service providers are a prime example. When one user sends data to another, the provider's equipment facilitates the transfer automatically. Does that mean that the provider is directly liable when the transmission happens to result in the "reproduc[tion]" of a copyrighted work? It does not. The provider's system is "totally indifferent to the material's content," whereas courts require "some aspect of volition" directed at the copyrighted material before direct liability may be imposed. The defendant may be held directly liable only if the defendant *itself* "trespassed on the exclusive domain of the copyright owner." Most of the time that issue will come down to who selects the copyrighted content: the defendant or its customers.
>
> . . . .
>
> The distinction between direct and secondary liability would collapse if there were not a clear rule for determining whether the *defendant* committed the infringing act. The volitional-conduct requirement supplies that rule; its purpose is not to excuse defendants from accountability, but to channel the claims against them into the correct analytical track.[72]

Thus, *Aereo* has caused some to wonder whether the majority had implicitly rejected a volitional act requirement for copyright infringement,[73]

[72] Id. at 2513–14 (Scalia, J., dissenting) (citations omitted) (first quoting 17 U.S.C. § 106(1) (2012); then quoting *CoStar Grp.*, 373 F.3d at 550–51; then quoting id. at 550). There is a conceptual connection between a volitional act requirement and certain forms of secondary liability in copyright law. In particular, the Supreme Court has held—with respect to secondary liability for a provider of peer-to-peer file-sharing software—that "one who distributes a device with the object of promoting its use to infringe copyright, as shown by clear expression or other affirmative steps taken to foster infringement, is liable for the resulting acts of infringement by third parties." Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913, 936–37 (2005). Just as the presence of volition indicates that a technology provider has gone beyond merely deploying its automated system to copy, inducement of third-party infringement indicates that a technology provider has gone beyond merely providing a system or device that can be used by others to infringe copyright.

[73] E.g., Bruce E. Boyden, *Aereo* and the Problem of Machine Volition, 2015 Mich. St. L. Rev. 485; Kyle A. Brown, Comment, Up in the *Aereo*: Did the Supreme Court Just Eliminate the Volitional Conduct Requirement for Direct Copyright Infringement?, 46 Seton Hall L. Rev. 243 (2015).

although the Second and Ninth Circuits have affirmed the requirement's continuing relevance.[74] Owing to the ongoing relevance of volition in copyright law, it is worth making sense of this requirement, to which the next section now turns.

D. *What Is Volition in Copyright Law?*

What exactly is this "volition" mental state required for copyright infringement liability? As this Essay noted earlier, volition might be understood as the mental state that causes willful actions.[75] In other words, the question of whether some event counts as volitional is the question of whether it is something genuinely willed or chosen by the so-called actor. The presence of a volitional mental state as a cause thus distinguishes involuntary bodily movements—such as those during a seizure—from voluntary ones.[76] With that distinction, a volition requirement coheres with the intuition that individuals should be held responsible for, and only for, that which was under their control.[77] As the Restatement (Second) of Torts explains, "Some outward manifestation of the defendant's will is necessary to the existence of an act which can subject him to liability."[78]

Note that copyright's volitional act requirement is asking for volition or control in something very specific: the production of the infringing copy itself. After all, technology providers have chosen—that is, willfully acted—in providing copy-making technologies, such that holding them responsible for resulting infringements would not constitute responsibility for something entirely out of their control.[79] Nonetheless, volitionally providing the technology is not sufficient for satisfying copyright law's volitional act requirement. Instead, copyright requires that the instance of infringing copying itself be volitional—or itself count as a willful action on the part of the technology provider—and that the infringing

[74] See BWP Media USA Inc. v. Polyvore, Inc., 922 F.3d 42, 49 (2d Cir. 2019) (per curiam) ("[W]e have reaffirmed post-*Aereo* . . . that '[v]olitional conduct is an important element of direct liability.'" (quoting EMI Christian Music Grp., Inc. v. MP3tunes, LLC, 844 F.3d 79, 96 (2d Cir. 2016))); Perfect 10, Inc. v. Giganews, Inc., 847 F.3d 657, 667 (9th Cir. 2017) (explaining that one element of a direct infringement claim is volitional conduct).

[75] See, e.g., Moore, Act and Crime, supra note 11 (canvassing and assessing different philosophical conceptions of volition); Robert Audi, Volition, Intention, and Responsibility, 142 U. Pa. L. Rev. 1675, 1680 (1994) ("Moore sees conflict as a pervasive element in our desire and belief systems. Action cannot occur without resolution of such conflicts; volition here plays the role of reconciler, or, at least, of referee.").

[76] See, e.g., Restatement (Second) of Torts § 2 cmt. a (Am. Law Inst. 1965) ("There cannot be an act without volition. Therefore, a contraction of a person's muscles which is purely a reaction to some outside force, such as a knee jerk . . . , are not acts of that person. . . . So too, movements of the body during sleep . . . are not acts.").

[77] Moore, Act and Crime, supra note 11, at 48.

[78] Restatement (Second) of Torts § 2 cmt. a.

[79] See Denicola, supra note 70, at 1265 (explaining that courts have found volition when defendants made a choice to deploy systems that made infringement possible).

conduct can be attributed to the provider rather than the technology user alone.[80] This is to say that copyright law asks for volition at a specific point on the causal chain: not simply the instance of providing copying technology but the particular instance of copying.[81] This requirement is plausibly motivated by the policy that it would be bad to hold technology providers responsible for all infringements resulting from their technologies—including ones proximately caused by someone else's actions—when these technologies are capable of value-adding, noninfringing uses and are therefore not ones that the law seeks to disincentivize entirely. For copyright, such technology providers thus must have volitionally "committed" the infringing action themselves, perhaps with opportunity to pause, evaluate, and then choose whether to proceed with the particular infringing action, in order to be held responsible for it.[82]

All in all, given that the legal attention to machine operations has been relatively extensive in the context of copyright's volitional act requirement, it provides a good test bed for exploring machine mental states more broadly across the law. For in copyright law, many courts have treated liability for human and mechanical, or automated, acts of copying dichotomously: Humans always have volition, even when they are copying subconsciously, whereas machines can—and, to some courts, always—lack volition, even when carrying out acts of copying for which they are centrally designed.[83] Indeed, the particularly strong language of

[80] See id. at 1272 (describing a hypothetical in which a customer uses a provider's machine to reproduce a copyrighted work to demonstrate that "[t]he volition requirement . . . defines the connection between the owner of a copying system and the copied work that is sufficient to justify attributing the copying of that work to the owner").

[81] An alternative way of describing copyright's volitional act requirement is that it requires that the actions of the technology provider be the proximate cause of the production of the copy for the technology provider to be liable for infringement. See, e.g., BWP Media USA Inc. v. Polyvore, Inc., 922 F.3d 42, 61–67 (2d Cir. 2019) (Newman, J., concurring in the result) ("Infringement is a tort . . . . 'Volition' . . . is best understood to mean a concept essentially reflecting tort law causation. . . . '[C]ausation,' in the context of copyright infringement, is tort law 'proximate cause,' rather than 'but for' causation."). Note that this interpretation of the volitional act requirement is ultimately equivalent to the interpretation we favor according to which volitions are the mental states causing actions, for it is asking whether the proximate cause of infringement is the action of the technology provider. Furthermore, what determines whether something counts as the technology provider's actions (rather than someone else's) is whether it is the result of the technology provider's (or its machine's) volitional mental state (which causes actions rather than mere movements).

[82] See, e.g., Moore, Act and Crime, supra note 11, at 111–65.

[83] Cf. James Grimmelmann, Copyright for Literate Robots, 101 Iowa L. Rev. 657, 657 (2016) ("Almost by accident, copyright law has concluded that it is for humans only: reading performed by computers doesn't count as infringement. Conceptually, this makes sense: Copyright's ideal of romantic readership involves humans writing for other humans."). Professor Matthew Sag has observed that whether machines or their owners are liable for copyright infringement ought to turn on whether the machines are copying

Cartoon Network seems to entail that if we imagine an (inefficient) internet whose computers—servers and all—are each replaced with a human given the task to copy received material and pass it on toward the specified destination, then this imagined internet would count as having volition under copyright law at each node, whereas the currently automated internet lacks it entirely.[84] This implication is notwithstanding the fact that both variations of the internet—by stipulation—would be functionally identical systems. But this thought experiment is reminiscent of those deployed by philosophers in their efforts to understand the nature of human minds and machines, to which we now turn.

III. The Philosophy of Mind and Machines

This Part surveys two of the most influential philosophical discussions on the mind—namely, John Searle's "Chinese Room" argument and David Chalmers's two concepts of mind—in order to explicate the important conceptual distinction between "conscious" and "functional" understandings of mental states. It then explains the implications of this philosophical distinction for the question of whether any of the law's mental state requirements, such as copyright law's volitional act requirement, can or should be satisfied by machines.

A. *John Searle and the "Chinese Room" Argument*

Philosophers of mind have long contemplated whether there is any fundamental difference between human and artificial minds. Perhaps the most well-known challenge to the possibility of computers with truly human-like mental states is John Searle's "Chinese Room" argument. This argument has shaped much of the course of philosophical thinking on these questions since its publication in 1980, spurring continuing debate about the possibility of so-called "strong" AI—purely computational systems that possess *conscious* mental states like those of humans—versus "weak" AI, which merely *functionally* simulates the human mind.[85] In particular, Searle asks us to consider the following thought experiment:

> Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese . . . . Now suppose further that after this first batch of Chinese writing I am given a second batch of

works for expressive or nonexpressive uses. See Matthew Sag, Copyright and Copy-Reliant Technology, 103 Nw. U. L. Rev. 1607, 1624–44 (2009).

[84] See Cartoon Network LP v. CSC Holdings, Inc., 536 F.3d 121, 131 (2d Cir. 2008) ("In determining who actually 'makes' a copy, a significant difference exists between making a request to a human employee, who then volitionally operates the copying system to make the copy, and issuing a command directly to a system, which automatically obeys commands and engages in no volitional conduct.").

[85] See Paul M. Churchland & Patricia Smith Churchland, Could a Machine Think?, Sci. Am., Jan. 1990, at 32, 32–34 (noting that "Searle's paper provoked a lively reaction from AI researchers, psychologists and philosophers alike").

> Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that "formal" means here is that I can identify the symbols entirely by their shapes. . . . Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols . . . that from the external point of view—that is, from the point of view of somebody outside the room in which I am locked—my answers to the questions are absolutely indistinguishable from those of native Chinese speakers.[86]

In other words, Searle asks us to imagine that he is performing computational operations on the Chinese characters in accordance with formal rules, thereby instantiating a computer program.[87] Although the program that he is operating has the same input–output structure as a human fluent in Chinese, such that it is computationally *equivalent* to a Chinese speaker, Searle argues that he—and the program—nonetheless lack the *conscious experience* of a Chinese speaker who genuinely understands the language.[88] In other words, he explains, there is a fundamental difference between what goes on in the Chinese Room and an alternative scenario in which Searle responds to English inputs with outputs on the basis of formal rules.[89] In the case of English, Searle is not solely *functionally* instantiating the English program but also consciously understands.[90] In the Chinese Room, however, he merely simulates a conscious Chinese speaker.[91]

Searle's thought experiment challenged both the view that it is possible for there to be an artificial system with conscious mental states resulting from purely computational processes[92] and the view that human

[86] John R. Searle, Minds, Brains, and Programs, 3 Behav. & Brain Sci. 417, 417–18 (1980). Note that Searle himself originally put forth the "Chinese Room" argument as a challenge to the possibility of computation-based *understanding* rather than *consciousness.* But some philosophers have subsequently interpreted the argument as actually challenging the possibility of an artificial computer *experiencing* understanding, which is ultimately the question of artificial consciousness. See, e.g., David Chalmers, The Conscious Mind: In Search of Fundamental Theory 322–23 (1996). For this Essay's purposes, we follow these philosophers' interpretation of Searle's argument. Nonetheless, we flag the alternative interpretation and note that the choice of interpretation ultimately has no bearing on this Essay's thesis.

[87] Searle, supra note 86, at 418.

[88] Id.

[89] Id.

[90] Id.

[91] Id.

[92] As Searle explains,

> Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena. No one would suppose that we could produce milk and sugar by running a

consciousness is itself simply the product of computation.[93] In other words, Searle argued, because the functional processes of computation cannot give rise to conscious mental states and because our human minds clearly possess such mental states, it cannot be the case that our human minds are solely instantiating a program.[94]

This argument triggered decades of discussion, including a slew of critical responses from philosophers, psychologists, and computer scientists. Some of these challenges reject Searle's conclusion about the Chinese Room, saying it in fact does experience understanding of Chinese, even if the person inside the room—who is only a part of the computational system—does not.[95] Others have said that even if the Chinese Room lacks such experience, this is only because it is running the wrong kind of program; if it were instead running, say, a program simulating all the intricacies of the human brain, then it would have the experience of a Chinese speaker.[96] But Searle himself has responded to these objections, even addressing many in his original paper;[97] and there thus remains a rift between those who find the Chinese Room to be compelling in showing that the human mind could not be a computer and those who regard the argument as fundamentally mistaken.

B. *David Chalmers and the Hard Problem of Consciousness*

Regardless of whether Searle's argument is successful, the conceptual distinction between conscious and functional properties of mental states—which is made particularly vivid by the Chinese Room argument—remains enormously important and is taken seriously by all such philosophers. Pointedly, even human mental states can be understood in

> computer simulation of the formal sequences in lactation and photosynthesis, but where the mind is concerned many people are willing to believe in such a miracle because of a deep and abiding dualism . . . .

Id. at 424.

[93] Id. ("Whatever it is that the brain does to produce intentionality, it cannot consist in instantiating a program since no program, by itself, is sufficient for intentionality.").

[94] Id.

[95] For example, Daniel Dennett posits that

> Searle, laboring in the Chinese Room, does not understand Chinese, but he is not alone in the room. There is also the System, . . . and it is to *that* self that we should attribute any understanding . . . .
>
> This reply to Searle's example is what he calls the systems reply. It has been the standard reply of people in AI from the earliest outings of his thought experiment.

Daniel C. Dennett, Consciousness Explained 439 (1991).

[96] See, e.g., Chalmers, supra note 86, at 323–25 (arguing that at least a system with the same functional organization or structure as a brain would mirror the "causal relations between neurons" and therefore have the same conscious properties); Churchland & Churchland, supra note 85, at 37 (arguing that a system mimicking a human brain might be conscious).

[97] Searle, supra note 86, at 419–22.

terms of either conscious or functional properties. David Chalmers famously made this point in *The Conscious Mind*, which articulated what he called the "hard problem" of consciousness.[98] As Chalmers explains, the term "conscious" might be understood as synonymous with "phenomenal," the idea being that if an entity is conscious, then there is something that it is like to be that being.[99] To illustrate this concept, consider the contrasting examples of a human and a thermometer. Although a human and a thermometer both possess functional attributes that enable them to detect heat, the human feels or experiences heat, whereas the thermometer does not.[100] This is the difference between beings—such as humans—that have the capacity for such subjective experiences and beings—such as thermometers—that do not: Only the former are conscious beings.

In his book, Chalmers demonstrates that individual human mental states can be analyzed either in terms of what he calls their *psychological* properties—their functional role in producing behavior, or what they do—or their *phenomenal* properties—their conscious quality, or how they feel.[101] That is, according to Chalmers, the functional and the conscious concepts of the mind are distinguishable, even with respect to the human mind.[102] Consider, for instance, Chalmers's example of the "pain" mental state.[103] Pains have conscious aspects: There is something it is like to be in pain (indeed, it is unpleasant).[104] But pains also have entirely functional properties, which specify their structural roles in causal systems. For example, a pain has the functional properties of typically being the product of some damage to one's body, leading to adverse reactions to the stimulus such as saying "ow," recoiling, and so forth.[105] Upon separating the two concepts of mind, Chalmers ultimately argues for the conceivability of an entity that possesses human mental states understood entirely in terms of their functional properties, but which nonetheless lacks any conscious experience of those states.[106] As he explains, the Easy Problem (despite being difficult in its own right) is the question of the precise functional nature of mental states;[107] the Hard Problem is the

[98] Chalmers, supra note 86, at xi–xii.

[99] Id. at 285–86. See generally Thomas Nagel, What Is It Like To Be a Bat?, 83 Phil. Rev. 435 (1974) (explicating the philosophical difficulties surrounding the concept of consciousness).

[100] At least, we plausibly suspect that it does not. An alternative view is offered by panpsychism, the idea that all objects possess conscious minds. See, e.g., Chalmers, supra note 86, at 297–301.

[101] Id. at 11.

[102] Id. at 17.

[103] Id.

[104] Id.

[105] Id.

[106] Id. at 17–18.

[107] Id. at xi–xii.

question of why or how certain beings—such as humans—also have conscious experience.[108]

This philosophical distinction between the conscious and functional properties of the mind has important implications for the law and its governance of machines. This is because, regardless of one's views on whether conscious AI is possible, most philosophers—including Searle—agree that machines (like the Chinese Room) can in principle replicate the functional properties of human minds.[109] Moreover, for each of the law's mental state requirements, it remains an open question whether the law ultimately seeks to track the conscious or functional properties of the states in question. Because the law has primarily been designed for human actors, for whom the conscious and the functional typically coincide, this is a question we have principally been able to avoid until now. But the increasing prevalence of ever-sophisticated machines requires us to take it seriously. If the law is concerned only with functional properties, then these properties could very well be possessed by the states of a nonhuman machine.[110] In other words, then, it is far from settled that all the law's mental state requirements should be satisfied only by conscious minds. The remainder of this Essay challenges this assumption, analyzing the case of the aforementioned volitional act requirement in copyright law.

IV. VOLITION AND AI: IS CONSCIOUSNESS RELEVANT?

This Part argues that the volition requirement in copyright law ultimately does not seem interested in tracking conscious properties of the human infringer but instead functional ones, which could in principle be possessed by a machine.

The earlier analysis of the purpose of copyright's volitional act requirement[111] still leaves open the question of whether such "volition" at the instance of infringement must be conscious rather than some functional analogue, or whether such a purely functional state of a machine can result in something that, at least for the law's purposes, should be regarded as a "willful action" on the part of the technology provider. In

[108] Id. at 4–5.

[109] See Searle, supra note 86, at 418 (granting that the Chinese Room is functionally "indistinguishable . . . from native Chinese speakers").

[110] According to one school of artificial intelligence, human-like intelligence in machines can emerge only from machines that are embodied with features that are human-like, such as the brain and eyes. See generally Rodney A. Brooks, Cambrian Intelligence: The Early History of the New AI (1999) (exploring how behavior-based robots can act in ways that appear intelligent); Andy Clark, Being There: Putting the Brain, Body, and World Together Again (A Bradford Book reprint ed. 1998) (1997) (theorizing how the brain is a controller for embodied activity, and deriving an action-oriented theory of the mind). To the extent that this school is correct, artificial intelligence will appear relatively human.

[111] See supra section II.D.

other words, given the law's concerns and that a business's human employees almost always count as "acting" on the part of the business for the law's purposes, is there a reason for thinking, as a categorical matter, that the business's nonconscious machines—no matter their functions—never could? We think the answer is "no."

Consider first the general question: When should *any* area of law require a conscious rather than so-called functional volition? One might argue that a being should be held legally responsible for itself—or as a conscious, autonomous agent—only if that being is genuinely conscious. But this thesis would certainly need to be defended, for it would depend on the purpose of liability in the particular legal domain. If the purpose is entirely to produce the proper incentives—the dominant American view of copyright[112]—then it is not clear why the actor being held responsible must have consciousness, rather than simply the right functional responses to such incentives. On the other hand, at least for some areas of law, one might have the view that legal responsibility is meant to track moral responsibility.[113] Such a theorist thus might argue that it is nonsensical to hold a nonconscious being morally responsible for its behavior, as such a being is not a moral agent. Underlying this claim is the premise that, for something to be a moral agent, it must have conscious experience. But even this supposition requires substantiation and is undoubtedly up for debate.[114] For instance, imagine a machine with all the functional properties of a human. Such a machine would thereby have the capacity for something functionally equivalent to moral deliberation and judgment, and for choosing an action on the basis of such judgment, all despite lacking any conscious experience of this process. We might thus wonder why these functional capacities are not themselves sufficient for moral agency, or why their conscious quality (or lack thereof) would be relevant to the question at all.

In any event, even if one embraces the view that a being must be conscious for it to be held legally responsible for itself, this ultimately does not pose a challenge for the suggestion—say, in the context of copyright law—that the mere functionality of a technology provider's machine could suffice for holding that provider responsible. This is because holding a human or business entity responsible for its machine (or,

[112] See supra text accompanying notes 30–37. See generally William M. Landes & Richard A. Posner, The Economic Structure of Intellectual Property Law (2003) (articulating and defending an economic understanding of the aims of intellectual property law).

[113] See, e.g., Michael S. Moore, Causation and Responsibility: An Essay in Law, Morals, and Metaphysics 4 (2009) ("[C]riminal and tort liability must track moral responsibility, because justice is achieved only if the morally responsible are held liable to punishment or tort damages.").

[114] See, e.g., S. Matthew Liao, The Basis of Human Moral Status, 7 J. Moral Phil. 159, 169 (2010) (arguing that the basis of human moral status is not the conscious properties of human beings but rather the fact that human beings possess the genetic basis for moral agency, and that nonhuman beings could also possess moral status).

indeed, its employee) does not seem to amount to treating said machine (or employee) as a conscious, autonomous agent; rather, it amounts to treating the human or business entity as responsible for the machine. In other words, whether or not machines themselves must have conscious mental states in order to be held responsible for their own so-called behavior, the question of whether a business entity can be held responsible for its machines—that is, whether these machines can be regarded as "acting" on said corporation's behalf—does not seem like it should turn on whether the machine in question is conscious.

Moreover, the idea that copyright's rules for infringement liability are ultimately unconcerned with consciousness is further suggested by the doctrine of subconscious copying, which has been widely criticized but nonetheless firmly remains a part of copyright law.[115] Recall that, under existing law, a human who subconsciously copies the work of another—that is, without any awareness that he or she is doing so—is still liable for copyright infringement.[116] Of course, the term "subconscious" as used in this doctrine is importantly different from the concept of phenomenal consciousness discussed earlier, for "subconscious" in the doctrine refers to the absence of awareness that an act of copying—rather than original creation—has occurred, whereas "unconscious" in the phenomenal sense refers to the absence of any phenomenal qualities whatsoever. But nonetheless, if copyright does not care about a potential infringer's awareness of their infringement, the question arises: Why think that it cares about the presence of any conscious awareness or experience whatsoever, even awareness of action? It is hard to see a reason to think it would. Indeed, when we imagine the case of a human employee operating a technology provider's copy machine—one whose mental states, we have seen, would always satisfy copyright's volition requirement—it seems plausible that this requirement might ultimately be interested in tracking what action the employee does and the function of their mind in facilitating this action, rather than their phenomenology while doing it.

V. A FUNCTIONAL UNDERSTANDING OF MACHINE VOLITION

The preceding discussion suggests that copyright's volition requirement may not demand consciousness and may instead be more concerned with functionality. The doctrinal upshot is that so-called "functional volition"—or functional properties that capture what the law

[115] See, e.g., Olufunmilayo B. Arewa, The Freedom to Copy: Copyright, Creation, and Context, 41 U.C. Davis L. Rev. 477, 531–39 (2007); Jessica Litman, Copyright as Myth, 53 U. Pitt. L. Rev. 235, 240 (1991); Wendy J. Gordon, Toward a Jurisprudence of Benefits: The Norms of Copyright and the Problem of Private Censorship, 57 U. Chi. L. Rev. 1009, 1029–31 (1990) (book review); Carissa L. Alden, Note, A Proposal to Replace the Subconscious Copying Doctrine, 29 Cardozo L. Rev. 1729, 1743–52 (2008).

[116] Supra text accompanying notes 38–40.

is ultimately interested in tracking here—may suffice for copyright, such that the operation of a machine could give rise to direct liability for the technology provider, rather than solely for the technology user. This upshot has undoubted practical significance: It makes a substantial difference to copyright owners who would otherwise be limited to attempting to hold only individual users directly liable, and it prevents technology providers from avoiding direct liability simply by replacing human employees with copy-making machines. But this framework raises the question of which functional properties the volitional act requirement might seek and what a machine would have to look like to possess them.

Consider the human reproducing copies of another's copyrighted work, whom copyright law says always possesses the requisite volition for infringement liability. Indeed, consider the human employee making copies of a protected work. Principal–agent liability would readily confer liability on the employer without a doubt as to the employee's volition.[117] Although such humans have the full range of the functional properties of a human mind, in which of these properties or capacities is the law ultimately interested in findings of infringement? Plausibly, it is not all of them, because the hypothetical copy-making humans do not use this full range of capacities. Instead, perhaps it is simply the humans' capacity to evaluate whether what they have been asked to copy is likely to be material within the realm of copyright subject matter—putting aside for the moment more complicated determinations such as the fair use defense to infringement, which is addressed shortly[118]—and to decline to make the copy on the basis of this determination. This functional capacity would align coherently with copyright law's stated aim to disincentivize third parties from copying protected materials in order to preserve the corresponding incentive that copyright offers to authors to create.[119] And more broadly, it would seem to comport with the volitional act requirement's general purpose of ensuring that the actor has had an opportunity to pause to evaluate whether to proceed in acting—and to decline to perform the action if she so chooses on the basis of this evaluation—before being held responsible for the action.[120]

This functional capacity seems relatively basic. Although it is not possessed by, say, a rudimentary copy machine—which is "compelled" to make copies upon the pressing of a button and therefore has no "choice" regardless of what is being copied—a more sophisticated computer could plausibly be designed to "choose" whether to make a copy, despite lacking the full range of human functional properties. In other words, such a computer—despite being functionally subhuman—would

[117] See infra text accompanying notes 128–132 (summarizing agency law).

[118] Infra text accompanying notes 123–124.

[119] See supra text accompanying notes 30–33.

[120] See supra section II.D.

be equivalent in all the ways copyright law cares about to a human operating a copy machine, who we already know is always "volitional" for purposes of copyright infringement.[121] For instance, bots fetching songs and software generating new art based on learning from existing artwork could readily possess volition in this sense of the word.[122]

On the other hand, perhaps the volitional act requirement seeks to track a more sophisticated functionality, such as the capacity to determine whether an instance of copying is likely to be fair use and choose to act on this determination. As Professor Dan Burk suggests, it is difficult—if not impossible—to devise algorithms that appropriately decide questions of fair use: "[T]he cost structure of algorithmic content policing has created a largely impersonal process, in which the context-specific factors that should be taken into account in fair use analysis are absent and go unconsidered."[123] In particular, Burk worries about the "human judgment" that must be baked into these systems ex ante or in evaluating machines' outputs ex post, such as a model of the markets for copyrighted works to assess the effect of a use on the market for a copyrighted work and the significance of the part of the work used.[124] Thus, if copyright law is interested in tracking the functional capacity to make plausible fair use determinations, then it seems that a functionally volitional machine remains far off.

Ultimately, this Essay does not aim to settle the question of the right functionality in which copyright law ought to be interested. Instead, it hopes to show that this is the type of question scholars and policymakers need to be asking, rather than simply assuming that machines can never be volitional as a matter of law.

Moreover, it must be emphasized that this conclusion is not merely one of philosophical interest. Rather, whether and when machines can possess the requisite volition to infringe copyright has great practical import. The precise contours of the volitional act requirement have implications for who is and is not directly accountable for the copying of protected material. For copyright law to accomplish its goals of encouraging the creation and dissemination of expressive works, it must provide

[121] Supra notes 64–68 and accompanying text.

[122] Supra text accompanying notes 19–21.

[123] Dan L. Burk, Algorithmic Fair Use, 86 U. Chi. L. Rev. 283, 290 (2019) (emphasis omitted).

[124] Id. at 296. Similarly difficult, as Sonia Katyal and Jason Schultz point out, are questions of which parts of a work are protectable as original and whether the author has expressly or implicitly licensed uses of the work. See Sonia K. Katyal & Jason M. Schultz, The Unending Search for the Optimal Infringement Filter, 112 Colum. L. Rev. Sidebar 83, 96–101 (2012), https://columbialawreview.org/wp-content/uploads/2016/05/Katyal-Schultz.pdf [https://perma.cc/V4EW-MNUT]. Other scholars are more supportive of the possibility of algorithmic copyright enforcement so long as the machine providers are transparent about and accountable for their substantive determinations. See Maayan Perel & Niva Elkin-Koren, Accountability in Algorithmic Copyright Enforcement, 19 Stan. Tech. L. Rev. 473, 477–78 (2016).

sufficient incentive to creators with copyright's exclusive rights and concomitant disincentive to third parties from infringing those rights by holding them liable for infringement.[125] Holding the providers of machines that act with the requisite volition directly liable for infringement thus plays an important role in doing just that. Indeed, even if there is also a technology user—a so-called customer—to hold accountable for infringing uses of a technology, this should not rule out holding qualifying technology providers liable for infringement as well. And given that technology users in these cases might be judgment proof while the technology provider frequently is not, the ability to hold the technology provider liable can have significant practical import. Moreover, because of its intricate connection to copyright policy, an inquiry into machine volition as a matter of direct liability will frequently be more pertinent and straightforward than an investigation of secondary liability, in light of the law's relatively mystifying standards for the latter.[126]

At this point, one might be concerned with the policy implications of a conclusion that machines can have functional mental states or that functionality is what matters for findings of copyright infringement. For instance, does this overly discourage innovation of more sophisticated technologies, ones which—unlike simple copy machines—possess functional volition, to the extent that technology providers will attempt to "design around" liability? Or should technology providers be required to employ functionally volitional machines? Perhaps it would be sufficient to require machines to flag certain (or all) material for review by a human—such as a lawyer—before copying it, and thereby introduce human volition at the instance of copying. Such a design would give the machine the ability to pause and evaluate before proceeding to copy protected material. But it might also incapacitate machines from automating many of the tasks we have come to expect from them, precisely as the *Netcom* court worried.[127]

Thus, the reader might wonder whether the forgoing discussion on human and machine volition should move us to reconsider the volitional act requirement itself. For instance, we might ask whether (on the one hand) a technology provider's volition in providing copying technology should be sufficient for liability rather than requiring volition at the

[125] See supra text accompanying notes 30–37.

[126] See Mark Bartholomew & John Tehranian, The Secret Life of Legal Doctrine: The Divergent Evolution of Secondary Liability in Trademark and Copyright Law, 21 Berkeley Tech. L.J. 1363, 1409–10 (2006) ("The difficulty of pursuing direct infringers has never served as a doctrinal basis for . . . secondary liability. Such reasoning undermines the stability of legal guidelines, rendering them unreliable . . . and erod[ing] the principled bases for secondary liability."); Lital Helman, Pull Too Hard and the Rope May Break: On the Secondary Liability of Technology Providers for Copyright Infringement, 19 Tex. Intell. Prop. L.J. 111, 123 (2010) (stating that Supreme Court case law on secondary liability for copyright infringement "may have actually sowed the seeds of confusion reflected in the area . . . to this day").

[127] Supra text accompanying note 60–61.

instance of copying itself, or whether (on the other hand) the doctrine of subconscious copying should be rejected. And, indeed, such skeptical musings are ones in which we ourselves are inclined to engage. But note that they bear on the question of whether the volitional act requirement is a good thing and not whether—given what it seems to be trying to do—machines of any kind would and should ever satisfy it. Questions of the latter sort, we have demonstrated, cannot be handled so indelicately as some courts seem to think, for the law—for better or for worse—very well might here be interested in tracking only functional properties. Thus, as we enter a world in which users ask bots to find particular songs online and software gathers existing artworks to learn to create new art, it is increasingly important that we address such questions with due care. Moreover, because copyright law's volitional act requirement has served only as a case study, note that—regardless of what should or does become of this particular requirement—the challenge posed by the rest of the law's countless mental state requirements remains. The presented framework offers a path forward in analyzing how to adapt these requirements to a technologically evolving world.

The preceding analysis has pushed back on the assumption that mental state requirements can be satisfied only by human minds, instead asking both whether the particular requirement in question is ultimately about conscious or functional properties, and what a machine would have to look like to possess the functional properties of interest. But this analysis has focused in detail on one example, considering the apparent aims of a specific mental state requirement in copyright law. The next Part thus moves to generalize a theory of machines' mental states.

VI. TOWARD A GENERAL THEORY

As noted at the outset, despite the analytical focus until now on copyright law, the point to be gleaned from the present Essay is ultimately general: In the case of each implicit or explicit mental state requirement in the law, legal scholars and policymakers will need to engage in a similar analysis while attending to the unique interests and values at stake with regard to that law, in order to determine whether consciousness or mere functionality is what matters.

Of course, even if machines can have functional mental states, they do not have money, rights, or status as legal persons (at least, for the time being). Thus, the consequence of our analysis is that—to the extent that machines might be understood as having mental states for the law's purposes—machines might cogently be understood as agents of the business principal that creates or deploys them, performing actions for which that principal can be held directly responsible.[128] As one scholar puts it, an

[128] See Anat Lior, The Artificial Intelligence Respondeat Superior Analogy 54 (unpublished manuscript) (on file with the *Columbia Law Review*). By contrast, if machines cannot be understood to possess mental states in the view of the law, it is likely that they

agent "functions as the principal's representative, as an extension of the principal, while retaining the agent's own separate legal personality."[129] Agency, as per the most recent Restatement of the Law on the topic, is "the fiduciary relationship that arises when one person (a 'principal') manifests assent to another person (an 'agent') that the agent shall act on the principal's behalf and subject to the principal's control, and the agent manifests assent or otherwise consents so to act."[130] An agent can act with actual or apparent authority from the principal vis-à-vis third parties.[131] When an agent does so, pursuant to principles of respondeat superior, the principal can be legally liable for the agent's actions.[132] Thus, by suggesting the possibility of machines with legally required mental states, we are ultimately suggesting that there are contexts in which such machines are (functionally) agents in all the ways that matter. For that reason, just as a business would be liable for the conduct of its human agents, a business that creates and deploys these machines should be liable as principals for the conduct of these machines.[133] The possibility of technology providers being directly liable for infringement by their functionally volitional copying technologies is only one example of how this might manifest.

would instead be perceived as instrumentalities of the businesses or individuals that create and deploy them. Cf. id. at 12 (discussing the possible analogy of artificially intelligent machines to property). The Restatement (Third) of Agency takes the position that computers circa 2006 cannot be agents on the ground that "[t]o be capable of acting as . . . an agent, it is necessary to be a person, which in this respect requires capacity to be the holder of legal rights and the object of legal duties." Restatement (Third) of Agency § 1.04, cmt. e (Am. Law Inst. 2006). According to the Restatement, "a computer program is not capable of acting as . . . an agent . . . . At present, computer programs are instrumentalities of the persons who use them." Id. In light of this Essay's analysis and the trajectory of artificial intelligence technology, this position may warrant reconsideration.

[129] Deborah A. DeMott, The Contours and Composition of Agency Doctrine: Perspectives from History and Theory on Inherent Agency Power, 2014 U. Ill. L. Rev. 1813, 1816.

[130] Restatement (Third) of Agency § 1.01.

[131] See id. §§ 2.01–2.02 (actual authority); id. § 2.03 (apparent authority).

[132] Id. § 2.04 (respondeat superior); id. § 2.06 (liability of undisclosed principal); id. §§ 7.03–7.08 (principal's liability for an agent's actions). According to agency principles, "[a]n agent is [also] subject to liability to a third party harmed by the agent's tortious conduct . . . although the actor acts as an agent or an employee, with actual or apparent authority, or within the scope of employment." Id. § 7.01. What this might mean with regard to artificially intelligent machines is beyond the scope of this Essay.

[133] Note that a complete defense of the idea that machines can and should sometimes be regarded as the agents of humans or corporations would also require an explication of what mental states (or other requirements) *humans* need to possess to count as a machine's principal. We set aside consideration of this question for future work. An important point to note, however, is that requirements for technology providers to be regarded as the principals of their functionally volitional machines are plausibly different from, and perhaps weaker than, what existing courts require of technology providers under their present (and, in our view, mistaken) understanding of the volitional act requirement.

To move toward that more general enquiry, one might start by considering some preliminary thoughts on two very different mental state requirements: namely, volitional act requirements in criminal law rather than copyright[134] and copyright's requirements for authorship rather than infringement.[135] One could coherently embrace the view that, although functionality is all that matters for volition in the context of copyright infringement, consciousness matters in both of these alternative legal contexts. For instance, one might argue that the punitive aims of criminal law ultimately require that those engaging in criminal conduct have a conscious experience of the actions in which they have engaged.[136] One might also argue that, because status as an author under copyright involves possessing rights of ownership in one's creative work, it ultimately requires personhood,[137] something which—the argument would go—requires possessing a conscious mind.[138] We neither defend nor reject either such line of argument, as to do so would involve distinct projects in their own right. Rather, we invoke these two additional contexts to illustrate the way such analyses might go and how they might differ from our primary example of copyright infringement, owing to the distinct aims and considerations at play in each context.

At this point, one might wonder about the availability of a general theory regarding when the law cares about conscious versus purely functional properties of mental states such that this framework need not be applied on a painstakingly case-by-case basis. Perhaps the search for such a theory is precisely where this Essay should lead future work. Nonetheless, as a preliminary hypothesis—one reacting to, and consistent with, the examples we have here discussed—it might be that the law is interested in conscious properties of mental states when it seeks to treat the actor in question as a rightsholder (such as in copyright authorship) or an autonomous and responsible agent (such as in criminal punishment). But in contexts in which the law is seeking simply to protect the rights or

[134] See Moore, Act and Crime, supra note 11, at 44–46.

[135] See, e.g., Shyamkrishna Balganesh, Causing Copyright, 117 Colum. L. Rev. 1, 11–47 (2017) [hereinafter Balganesh, Causing Copyright] (defending and analyzing the idea of "authorial causation" as a requirement for copyrightability).

[136] See Samuel W. Buell & Lisa Kern Griffin, On the Mental State of Consciousness of Wrongdoing, 75 Law & Contemp. Probs., no. 2, 2012, at 133, 139–44 (exploring how blameworthiness can justify a requirement of conscious awareness of wrongdoing); cf. Shapira-Ettinger, supra note 16, at 2578 ("A normative theory [of guilt in criminal law] stands in contrast to the dominant psychological theory of guilt . . . prevailing . . . in legal systems today. The focus of the psychological approach to guilt is on . . . the internal state of mind that reflects the kind of consciousness with which one acts.").

[137] Balganesh, Causing Copyright, supra note 135, at 27 ("Given that authorship was invariably tied to ownership and the assertion of legal rights, it made little sense to speak of nonhuman authorship.").

[138] See generally Christopher Buccafusco, A Theory of Copyright Authorship, 102 Va. L. Rev. 1229 (2016) (defending a theory of authorship that requires intent: namely, the intention to produce mental effects in an audience).

interests of others from the actor (such as copyright infringement), functionality might be all that matters.[139] A thorough exploration or defense of this preliminary hypothesis is reserved for future work. But we hope this Essay has impressed the need to engage in such explorations and to wrestle with the fundamental questions surrounding the law's aims, in order to adapt the law to an increasingly machine-filled world.

Conclusion

In sum, it is a mistake to assume that machines can or should never satisfy implicit or explicit mental state requirements, entirely by virtue of the fact that they are machines. The law is not always or necessarily concerned with the existence of conscious experience or even with the full range of human-level functionalities. Instead, it will always be a substantive question what the law's various mental state requirements are aiming to track, one which depends on the interests and values at stake in the particular legal domain. It follows from this that, in adapting the law to a world with increasingly sophisticated technologies replacing the actions of humans, the challenge for the law is not that mental state requirements exist. Rather, it is that scholars and policymakers must start asking the normative questions of what such requirements are designed to achieve and therefore what relevant mental states must be.

[139] Thanks to Erick Sam for suggesting this hypothesis in conversation.

DATA-INFORMED DUTIES IN AI DEVELOPMENT

Frank Pasquale[*]

Law should help direct—and not merely constrain—the development of artificial intelligence (AI). One path to influence is the development of standards of care both supplemented and informed by rigorous regulatory guidance. Such standards are particularly important given the potential for inaccurate and inappropriate data to contaminate machine learning. Firms relying on faulty data can be required to compensate those harmed by that data use—and should be subject to punitive damages when such use is repeated or willful. Regulatory standards for data collection, analysis, use, and stewardship can inform and complement generalist judges. Such regulation will not only provide guidance to industry to help it avoid preventable accidents. It will also assist a judiciary that is increasingly called upon to develop common law in response to legal disputes arising out of the deployment of AI.

INTRODUCTION

Corporations will increasingly attempt to substitute artificial intelligence (AI) and robotics for human labor.[1] This evolution will create novel situations for tort law to address. However, tort will only be one of several types of law at play in the deployment of AI. Regulators will try to forestall problems by developing licensing regimes and product standards. Corporate lawyers will attempt to deflect liability via contractual arrangements.[2] The interplay of tort, contract, and regulation will not

[*] Piper & Marbury Professor of Law, University of Maryland. I would like to thank the *Columbia Law Review* staff for careful editing of the piece, as well as those commenting on the piece at the *Columbia Law Review* Symposium "Common Law for the Age of AI" (Colleen Chien and Olga Russakovsky). I also wish to thank attendees at the University of Melbourne's "Digital Citizen" conference and participants in a workshop at Data & Society entitled "Algorithms on the Shop Floor." I am particularly grateful to Taylor M. Cruz, Madeleine Elish, and Emanuel Moss for their comments at the Data & Society Workshop. Colleagues at the University of Maryland have also been generous with their time and expertise. Responsibility for errors or omissions, of course, rests with me.

[1] See, e.g., James Manyika, Susan Lund, Michael Chui, Jacques Bughin, Jonathan Woetzel, Parul Batra, Ryan Ko & Saurabh Sanghvi, McKinsey & Co., Jobs Lost, Jobs Gained: What the Future of Work Will Mean for Jobs, Skills, and Wages 1 (2017), https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Future%20of%20Organization/What%20the%20future%20of%20work%20will%20mean%20for%20jobs%20skills%20and%20wages/MGI-Jobs-Lost-Jobs-Gained-Report-December-6-2017.ashx [https://perma.cc/XCF4-JJPC] (describing the far-reaching impact that automation will have on the global workforce).

[2] This is already a common practice in the digital economy. See, e.g., Timothy J. Calloway, Cloud Computing, Clickwrap Agreements, and Limitation on Liability Clauses: A Perfect Storm?, 11 Duke L. & Tech. Rev. 163, 173 (2012) (describing a proliferation of limitation of liability clauses); Aaron T. Chiu, Note, Irrationally Bound: Terms of Use

just allocate responsibility ex post, spreading the costs of accidents among those developing and deploying AI, their insurers, and those they harm. This matrix of legal rules will also deeply influence the development of AI, including the industrial organization of firms, and capital's and labor's relative share of productivity and knowledge gains.

Despite these ongoing efforts to anticipate the risks of innovation, there is grave danger that AI will become one more tool for deflecting liability, like the shell companies that now obscure and absorb the blame for much commercial malfeasance.[3] The perfect technology of irresponsible profit would be a robot capable of earning funds for a firm, while taking on the regulatory, compliance, and legal burden traditionally shouldered by the firm itself. Any proposal to grant AI "personhood" should be considered in this light.[4] Moreover, both judges and regulators should begin to draw red lines of responsibility and attribution now, while the technology is still nascent.[5]

Licenses and the Breakdown of Consumer Rationality in the Market for Social Network Sites, 21 S. Cal. Interdisc. L.J. 167, 195 (2011) (describing the use of "disclaimers of liability" in social media network use agreements). For a practical example of how contracts are used to deflect, allocate, or redirect liability in the construction industry, see generally Patricia D. Galloway, The Art of Allocating Risk in an EPC Contract to Minimize Disputes, Construction Law., Fall 2018, at 26 (discussing risk allocation in engineering, procurement, and construction (EPC) contracts). In the health care context, "hold harmless" clauses can deflect liability from software providers. See Ross Koppel, Uses of the Legal System that Attenuate Patient Safety, 68 DePaul L. Rev. 273, 275–76 ("The 'hold harmless' clause in EHR [Electronic Health Record] contracts functions to prevent vendors from being held responsible for errors in their software even if the vendor has been repeatedly informed of the problem and even if the problem causes harm or death to patients.").

[3] As leading AI ethics expert Joanna Bryson has explained:

> Many of the problems we have in the world today come from people trying to evade the accountability of democracies and regulatory bodies. And AI would be the ultimate shell company. If AI is human-like, the argument goes, then you can use human justice on it. But that's just false. You can't even use human justice against shell companies. And there's no way to build AI that can actually care about avoiding corruption or obeying the law. So it would be a complete mistake—a huge legal, moral and political hazard—to grant rights to AI.

Fraser Myers, AI: Inhuman After All?, Spiked-Online (June 14, 2019), https://www.spiked-online.com/2019/06/14/ai-inhuman-after-all/ [https://perma.cc/A26G-YEX4] (conducting an interview with Bryson).

[4] See Joanna J. Bryson, Mihailis E. Diamantis & Thomas D. Grant, Of, for, and by the People: The Legal Lacuna of Synthetic Persons, 25 Artificial Intelligence & L. 273, 273 (2017) ("We review the utility and history of legal fictions of personhood, discussing salient precedents where such fictions resulted in abuse or incoherence. We conclude that difficulties in holding 'electronic persons' accountable when they violate the rights of others outweigh the . . . moral interests that AI legal personhood might protect.").

[5] Some may argue it is already too late, thanks to the power of leading firms in the AI space. However, there have been many recent efforts to understand and curb the worst effects of such firms. The U.S. government has demonstrated an interest in keeping large tech companies in line. For example, Facebook is currently facing a $5 billion fine from the FTC, a $100 million fine from the SEC, and an FTC antitrust investigation. Ian Sherr,

It may seem difficult to draw such red lines, because both journalists and technologists can present AI as a technological development that exceeds the control or understanding of those developing it.[6] However, the suite of statistical methods at the core of technologies now hailed as AI has undergone evolution, not revolution.[7] Large new sources of data have enhanced its scope of application, as well as technologists' ambitions.[8] But the same types of doctrines applied to computational sensing, prediction, and actuation in the past can also inform the near future of AI advance.[9]

A company deploying AI can fail in many of the same ways as a firm using older, less avant-garde machines or software. This Essay focuses on one particular type of failing that can lead to harm: the use of inaccurate or inappropriate data in training sets for machine learning. Firms using faulty data can be required to compensate those harmed by that data use—and should be subject to punitive damages when such faulty data

Facebook's $5 Billion FTC Fine Is Just the Start of Its Problems, CNET (July 25, 2019), https://www.cnet.com/news/facebooks-5-billion-ftc-fine-is-just-the-start-of-its-problems/ (on file with the *Columbia Law Review*). The Department of Justice is also reviewing tech companies for antitrust issues. Brent Kendall, Justice Department to Open Broad, New Antitrust Review of Big Tech Companies, Wall St. J. (July 23, 2019), https://www.wsj.com/articles/justice-department-to-open-broad-new-antitrust-review-of-big-tech-companies-11563914235 (on file with the *Columbia Law Review*). In response, tech companies, such as Facebook and Google, have expanded their lobbying capacity. See Cecilia Kang & Kenneth P. Vogel, Tech Giants Amass a Lobbying Army for an Epic Washington Battle, N.Y. Times (June 5, 2019), https://www.nytimes.com/2019/06/05/us/politics/amazon-apple-facebook-google-lobbying.html (on file with the *Columbia Law Review*).

[6] See, e.g., Will Knight, The Dark Secret at the Heart of AI, MIT Tech. Rev. (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ [https://perma.cc/V3LF-KBLD] (describing Nvidia's experimental autonomous car as having a "mysterious mind" unable to be understood by those designing it); David Weinberger, Our Machines Now Have Knowledge We'll Never Understand, WIRED (Apr. 18, 2017), https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/ [https://perma.cc/FW94-L2BE] ("This infusion of alien intelligence is bringing into question the assumptions embedded in our long Western tradition.").

[7] See, e.g., Best Practice AI, Evolution, Not Revolution: What the Bestpractice.ai Library Tells Us About the State of AI (Part 1), Medium (Sept. 17, 2018), https://medium.com/@bestpracticeAI/evolution-not-revolution-what-the-bestpractice-ai-library-tells-us-about-the-state-of-ai-part-1-f488b29add0b [https://perma.cc/VB86-544K] (describing findings from the development of Bestpractice.ai, a library of AI use cases and case studies).

[8] See generally Yoav Shoham, Raymond Perrault, Erik Brynjolfsson, Jack Clark, James Manyika, Juan Carlos Niebles, Terah Lyons, John Etchemendy, Barbara Grosz & Zoe Bauer, Artificial Intelligence Index: 2018 Annual Report (2018), http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf [https://perma.cc/3PWE-B7Z8] (presenting data suggesting that the number of patents and academic papers involving AI, among other metrics, have grown rapidly).

[9] Notable recent U.S. work in this vein includes Bryan Casey, Robot Ipsa Loquitur, Geo. L.J. (forthcoming 2019) (manuscript at 8–11), https://ssrn.com/abstract=3327673 (on file with the *Columbia Law Review*) (arguing that extant forms of liability should apply to robotics and thus many of the forms of AI that comprise the information processing of such robotics and can address many of the problems posed by such technology).

collection, analysis, and use is repeated or willful. Skeptics may worry that judges and juries are ill-equipped to make determinations about appropriate data collection, analysis, and use. However, they need not act alone—regulation of data collection, analysis, and use already exists in other contexts.[10] Such regulation not only provides guidance to industry to help it avoid preventable accidents and other torts. It also assists judges assessing standards of care for the deployment of emerging technologies. The interplay of federal regulation of health data with state tort suits for breach of confidentiality is instructive here: Egregious failures by firms can not only spark tort liability but also catalyze commitments to regulation to prevent the problems that sparked that liability, which in turn should promote progress toward higher standards of care.[11]

Preserving the complementarity of tort law and regulation in this way (rather than opting to radically diminish the role of either of these modalities of social order, as premature preemption or deregulation might do) is wise for several reasons. First, this hybrid model expands opportunities for those harmed by new technologies to demand accountability.[12] Second, the political economy of automation will only fairly distribute expertise and power if law and policy create ongoing incentives for individuals to both understand and control the AI supply chain and AI's implementation. Judges, lawmakers, and advocates must avoid developing legal and regulatory systems that merely deflect responsibility, rather than cultivate it, lest large firms exploit well-established power imbalances to burden consumers and workers with predictable harms arising out of faulty data.

I. Problems Caused by Inaccurate and Inappropriate Data

At its best, tort law rectifies wrongs (retrospectively) and enables persons to better plan their lives (prospectively).[13] This Part discusses some classic wrongs addressed by tort law and how the rise of AI, including the rhetoric surrounding it, may unnecessarily complicate adjudication arising out of them. To clarify some critical issues of duty and causation, litigants and courts should begin to focus on questions of inaccurate and inappropriate data, given the importance of data to the development of AI.

The duties of care prescribed by tort are reassuring aspects of a just social order. If a person is injured in a car accident by a negligent driver, courts should ensure some compensatory (and potentially punitive) damages payable by the tortfeasor (or their insurer) to ensure, as well as

[10] See infra Part II.

[11] See infra Part II.

[12] See Mary L. Lyndon, Tort Law and Technology, 12 Yale J. on Reg. 137, 143 (1995) ("The liability system supplements regulation.").

[13] See Melvin Aron Eisenberg, The Nature of the Common Law 4–5, 47–48 (1st ed. 1988).

possible, that the plaintiff is returned to the state of financial and physical health they would have enjoyed before the accident.[14] In the medical context, malpractice law is designed to give patients reassurance that if their physician falls below a standard of care, a penalty will be imposed and some portion of it dedicated to the recovery of the patient.[15]

The machines used by drivers and doctors are also subject to forms of tort liability: for example, in case they are negligently manufactured or defective by design.[16] These doctrines should have renewed relevance as new technologies of diagnosis and prediction arise in both general and specialty medical care. While AI applications promise many advances, they also create new risks.

Consider the rise of clinical decision support software for dermatologists. As the *Atlantic* recently reported, "A study that tested machine-learning software in dermatology, conducted by a group of researchers primarily out of Germany, found that 'deep-learning convolutional neural networks,' or CNN, detected potentially cancerous skin lesions better than the 58 dermatologists included in the study group."[17] To the extent such AI is continually validated, it may well become part of the standard of care for many tasks now performed by physicians.[18] However, the mere fact that a technology is better *in general* does not mean that it is optimal for all cases. In the case of facial recognition, there is a well-documented failure of AI systems to recognize the faces of persons of color, relative to its ability to recognize white persons' faces.[19] Many scholars have raised similar concerns with respect to racial disparities in health care in the

[14] Cf. Stuart M. Speiser, Charles F. Krause, Alfred W. Gans & Monique C. M. Leahy, American Law of Torts § 8:1 (Mar. 2019 Update) (describing the types of redress available to plaintiffs in a tort action).

[15] Alex Stein, Toward a Theory of Medical Malpractice, 97 Iowa L. Rev. 1201, 1203, 1209 (2012) ("Under the prevalent doctrine, a doctor commits malpractice when he treats a patient in a way that deviates from the norms established by the medical profession. The applicable norms flow from the accepted, or customary, medical practice: the ways in which similarly situated medical practitioners treat patients."). I introduce the topic with examples from transport and health in part because these fields are among the most affected, or likely to be affected, by advances in AI.

[16] See, e.g., Adams v. Toyota Motor Corp., 867 F.3d 903, 917 (8th Cir. 2017) (concluding that the evidence supported a jury verdict finding the manufacturer liable for deaths and injuries of persons involved in the collision in family members' products liability action based on a design defect).

[17] Angela Lashbrook, AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind, Atlantic (Aug. 16, 2018), https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/ [https://perma.cc/NLC2-VCFS].

[18] A. Michael Froomkin, Ian Kerr & Joelle Pineau, When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning, 61 Ariz. L. Rev. 33, 35, 61–63 (2019).

[19] See Tim Simonite, The Best Algorithms Struggle to Recognize Black Faces Equally, WIRED (July 22, 2019), https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/ [https://perma.cc/QQ4J-XBMB].

United States.[20] Physicians and computer scientists are already concerned that skin anomaly–detecting software may fail to work for African Americans and other minority groups in the United States as well as it does for white patients.[21]

Such problems are not new. In many cases, AI is little more than a better-marketed form of statistics, and consulting statistics has long been a part of medical practice.[22] AI is but one of many steps taken over the past two decades to modernize medicine with a more extensive evidence base.[23] Commentators have seized on predictive analytics, big data, artificial intelligence, machine learning, and deep learning as master metaphors for optimizing system performance.[24] Thus literature on each of these areas can illuminate the path forward for identifying problematic data in AI. Moreover, an emerging literature on the limits of AI (including lack of reproducibility, narrow validity, overblown claims, and opaque data) should also inform legal standards.[25]

[20] See, e.g., Dorothy Roberts, Fatal Invention: How Science, Politics, and Big Business Re-Create Race in the Twenty-First Century 81–103 (2012). See generally Dayna Bowen Matthew, Just Medicine: A Cure for Racial Inequality in American Health Care (2015) (examining racial health disparities through the lens of implicit bias).

[21] Adewole S. Adamson & Avery Smith, Machine Learning and Health Care Disparities in Dermatology, 154 JAMA Dermatology 1247, 1247 (2018). A cognate problem has arisen in genomics. See Eric Topol & Kai Fu Lee, It Takes a Planet, 37 Nature Biotechnology 858, 859 (2019) ("AI algorithmic development and validation requires diverse and massive datasets. There is little evidence for saturation but plenty of examples of misleading outputs when the data inputs are limited or venue specific.").

[22] See Meredith Broussard, Artificial Unintelligence: How Computers Misunderstand the World 32 (2018) ("Narrow AI is statistics on steroids.").

[23] See Inst. of Med. Roundtable on Evidence-Based Medicine, The Learning Healthcare System: Workshop Summary 81 (LeighAnne Olsen, Dara Aisner & J. Michael McGinnis eds., 2007), https://www.ncbi.nlm.nih.gov/books/n/nap11903/pdf/ [https://perma.cc/3VYA-M3S4] ("An essential component of the learning healthcare system is the capacity to continually improve approaches to gathering and evaluating evidence, taking advantage of new tools and methods.").

[24] See, e.g., Martin Ford, Architects of Intelligence 4 (2018) (describing deep learning); Viktor Mayer-Schönberger & Kenneth Cukier, Big Data: A Revolution that Will Transform How We Live, Work, and Think 7 (2013) ("Big data marks the beginning of a major transformation."); Nils J. Nilsson, The Quest for Artificial Intelligence 415 (2010) (describing reinforcement learning).

[25] Eric Topol, Deep Medicine 94 (2019) (citing concerns about "cherry-picking results or lack of reproducibility"); danah boyd & Kate Crawford, Critical Questions for Big Data, 15 Info., Comm. & Soc'y 662, 666–68 (2012) (describing how claims of objectivity and accuracy in big data can be misleading); Matthew Zook, Solon Barocas, danah boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A. Koenig, Jacob Metcalf, Arvind Narayanan, Alondra Nelson & Frank Pasquale, Editorial, Ten Simple Rules for Responsible Big Data Research, PLOS Computational Biology, Mar. 30, 2017, at 1, 2, https://journals.plos.org/ploscompbiol/article/file?id=10.1371/journal.pcbi.1005399&type=printable (on file with the *Columbia Law Review*) (identifying similar limits).

A. *Inaccurate Data*

In 2012, law professor Sharona Hoffman and computer scientist Andy Podgurski analyzed some common problems in then-emerging uses of big data in healthcare.[26] A great deal of the data that is now set to inform AI applications in healthcare is "generally observational, not experimental, and hence treatments and exposures are not assigned randomly. This makes it much more difficult to ensure that causal inferences are not distorted by systematic biases."[27] Dr. Dhruv Kullar gives a good example of the dangers of these dynamics:

> In medicine, unchecked A.I. could create self-fulfilling prophesies that confirm our preexisting biases, especially when used for conditions with complex trade-offs and high degrees of uncertainty. If, for example, poorer patients do worse after organ transplantation or after receiving chemotherapy for end-stage cancer, machine learning algorithms may conclude such patients are less likely to benefit from further treatment—and recommend against it.[28]

There are several problems with basing treatment on socioeconomic status. A skilled medical practitioner should be interested in *why* poorer patients are doing worse, not simply *that* they are.[29] Perhaps they have a harder time accessing follow-up care or healthy food. The proper response in that case is not to allow poverty to reduce the priority of a patient for a transplant. Rather, it is to invest in transportation, nutritional advice and subsidies, and other social supports that will promote a more

[26] Sharona Hoffman & Andy Podgurski, Big Bad Data: Law, Public Health, and Biomedical Databases, 41 J.L. Med. & Ethics (Spring Supp.) 56, 56 (2013).

[27] Id. at 57.

[28] Dhruv Khullar, Opinion, A.I. Could Worsen Health Disparities, N.Y. Times (Jan. 31, 2019), https://www.nytimes.com/2019/01/31/opinion/ai-bias-healthcare.html (on file with the *Columbia Law Review*). As Judea Pearl and Dana MacKenzie have shown, adding accounts of causation via diagrams and other intuitive explanatory tools can help professionals avoid such mistakes. Judea Pearl & Dana MacKenzie, The Book of Why: The New Science of Cause and Effect 13, 39–46 (2018). This is one reason why the European Union has adopted rules designed to promote explainable AI. See High-Level Expert Grp. on Artificial Intelligence, European Comm'n, Ethics Guidelines for Trustworthy AI 21–22 (2019), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 [https://perma.cc/7BKM-VDHP].

[29] As Hoffman and Podgurski put it:

> Confounding bias is a systematic error that occurs because there exists a common cause of the treatment/exposure variable and the outcome variable. For example, socioeconomic factors may be confounders because low income may cause individuals to choose sub-optimal, inexpensive treatments and may also separately lead to deteriorated health because of stress or poor nutrition. A failure to account for socioeconomic status may thus skew study results.

Hoffman & Podgurski, supra note 26, at 58 (footnote omitted).

successful transplant.[30] The main problem with the example Khullar gives is that poverty itself is not a direct cause of the bad medical outcomes.[31] Rather, there are intervening causes. AI scholars have long addressed this problem. For example, Judea Pearl and Dana MacKenzie have insisted that a knowledge of causation—*how* an alleged effect generates a cause—is crucial to genuine advances in AI.[32]

Moreover, even if it turns out that, ceteris paribus, poorer individuals simply do not do as well as others after transplants (surviving a shorter period of time, or with worse comorbidities and sequelae of the procedure), that fact alone would not dictate any particular change in their priority for organ transplantation. Society may decide that a thoroughgoing equality of access is the proper baseline for access to scarce organs, even if such allocation rules fail to maximize quality-adjusted life years (QALYs) or similar outcome metrics.[33]

Hoffman and Podgurski also point out the inadequacies of some data, especially those captured on the fly by doctors and nurses who already have more than enough to do on their shifts.[34] Electronic health record (EHR) systems may use different abbreviations: "Different systems may use different terminology to mean the same thing or the same terminology to mean different things. For example, the abbreviation 'MS' can mean 'mitral stenosis,' 'multiple sclerosis,' 'morphine sulfate,' or 'magnesium sulfate.'"[35] At present, the job of correcting (or throwing out) bad data, as well as related tasks of semantic harmonization and standardization, is often treated as secondary or menial.[36] But at a certain level of prevalence, such errors could be disastrous. Researchers must take into account measurement biases, which "are generated by errors in measurement and data collection resulting from faulty equipment or software or from human error."[37] Data are always socially shaped.[38] To

[30] See, e.g., Mary Simmerling, Beyond Scarcity: Poverty as a Contraindication for Organ Transplantation, 9 AMA J. Ethics 441, 442–44 (2007) (examining the financial burdens of post-transplant medications on the uninsured, the underinsured, and the poor).

[31] See Khullar, supra note 28.

[32] Pearl & MacKenzie, supra note 28, at 1–21. See generally Judea Pearl, Causal Inference in Statistics: An Overview, 3 Stat. Surv. 96 (2009) (discussing advances in statistical research that facilitate solving causal questions).

[33] See Jon Elster, Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens 22, 35–38 (1992) (discussing diverse normative bases for allocation decisions).

[34] Hoffman & Podgurski, supra note 26, at 57.

[35] Id. at 57.

[36] See Lilly Irani, Justice for "Data Janitors," Pub. Books (Jan. 15, 2015), https://www.publicbooks.org/justice-for-data-janitors/ [https://perma.cc/JLY6-Y6URt] (describing the work done by human "data janitors" to parse information that artificial intelligence systems are not capable of differentiating).

[37] Hoffman & Podgurski, supra note 26, at 58.

avoid troubling outcomes downstream, law must incentivize health care providers to ensure that data providers take the time and effort necessary to address well-known biases and shortcomings of data.

In the case of automobiles, similar problems may emerge. There may be certain individuals that a collision avoidance detection system is less likely to identify as persons.[39] Operators of autonomous cars may deploy humans as a backup, to ensure the data a car is reacting to are accurate, but even such a failsafe may itself be blameworthy if improperly applied. Human–computer interaction research has revealed that such "backup" roles are notoriously difficult to perform well, particularly in contexts in which attention is only required rarely and sporadically.[40]

B. *Inappropriate Data*

While earlier versions of AI, such as expert systems, were primarily rules based, data drives modern machine learning.[41] As recent controversies over predictive policing have shown, data can be unfairly unrepresentative: If minority neighborhoods have been overpoliced in the past, more crime will have been found in them than would be found in other neighborhoods, ceteris paribus.[42] Similarly, a firm that primarily hired

[38] See Lisa Gitelman & Virginia Jackson, Introduction, *in* "Raw Data" Is an Oxymoron 1, 2–6 (Lisa Gitelman ed., 2013) (arguing that data are not inherently neutral but rather constructed and gathered in ways that are shaped by academic disciplines). See generally Taylor M. Cruz, The Making of a Population: Challenges, Implications, and Consequences of the Quantification of Social Difference, 174 Soc. Sci. & Med. 79 (2017) (discussing how the process of gathering population data imposes implicit categorical assumptions on a heterogenous population).

[39] Benjamin Wilson, Judy Hoffman & Jamie Morgenstern, Predictive Inequity in Object Detection, arXiv (Feb. 21, 2019), https://arxiv.org/pdf/1902.11097.pdf (on file with the *Columbia Law Review*) (identifying potential for object detection technology to fail to detect people with darker skin tones).

[40] See, e.g., David A. Mindell, Our Robots, Ourselves 201–02 (2015) (describing the difficulties and failures associated with human operators serving as a backup in the event of failures by AI-driven systems such as autonomous vehicles); Madeleine Clare Elish, Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction, 5 Engaging Sci., Tech., & Soc'y 40, 52–55 (2019) (noting the difficulty of distributing responsibility and agency between a self-driving car and its safety driver).

[41] Pedro Domingos, The Master Algorithm 7 (2015).

[42] See Angèle Cristin, Predictive Algorithms and Criminal Sentencing, *in* The Decisionist Imagination: Sovereignty, Social Science, and Democracy in the 20th Century 272, 279–80 (2019) ("When predictive algorithms identify 'hot spot' crime zones (usually low-income African American neighborhoods), policemen are more likely to patrol in these neighborhoods and arrest people who will later be convicted. . . . This data will later be entered into the algorithm, thus producing a feedback loop.").

male managers in the past may end up developing AI hiring mechanisms that correlate success with gender, as opposed to actual job performance.[43]

Activists and authors are now exposing numerous examples of problematic data sets. For example, Caroline Criado Perez has explained how data sets often do not adequately represent women, with very troubling results.[44] In too much medical research and pedagogy, for instance, maleness is assumed as a default. As Perez asks, "There are still vast medical gender data gaps to be filled in, but the past twenty years have demonstrably proven that women are not just smaller men: male and female bodies differ down to a cellular level. So why aren't we teaching this?"[45]

Data may also be illegally obtained and therefore inappropriate for use. For example, an AI hiring algorithm might incorporate breached medical records that help it predict an applicant's health issues. Even if such health issues would impair the applicant's job performance, this data use is suspect. Thanks to trade secrecy, it may be difficult to detect or litigate.[46] Nevertheless, litigants are becoming increasingly sophisticated at unearthing the true bases of decisionmaking, and no firm should be entitled to hide the use of illegally obtained data.[47]

[43] See, e.g., Gideon Mann & Cathy O'Neil, Hiring Algorithms Are Not Neutral, Harv. Bus. Rev. (Dec. 9, 2016), https://hbr.org/2016/12/hiring-algorithms-are-not-neutral [https://perma.cc/BA6V-492D] ("When humans build algorithmic screening software, they may unintentionally determine which applicants will be selected or rejected based on outdated information—going back to a time when there were fewer women in the workforce, for example—leading to a legally and morally unacceptable result."); see also Miranda Bogen & Aaron Reike, Upturn, Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias 8–9 (2018), https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20–%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms,%20Equity%20and%20Bias.pdf [https://perma.cc/5T6U-4QL4] (describing examples of potential bias in predictive hiring tools).

[44] See generally Caroline Criado Perez, Invisible Women: Data Bias in a World Designed for Men (2019) (examining the "gender data gap").

[45] Id. at 199.

[46] See Sonia K. Katyal, The Paradox of Source Code Secrecy, 104 Cornell L. Rev. (forthcoming 2019) (manuscript at 104–05) (footnote omitted), https://ssrn.com/abstract=3409578 (on file with the *Columbia Law Review*) ("At their core, these automated systems often implicate central issues of due process, criminal (and civil) justice, and equal protection. Yet, because their inner workings are often protected as trade secrets, they can remain entirely free from public scrutiny."); Frank Pasquale, Digital Star Chamber, Aeon (Aug. 18, 2015), https://aeon.co/essays/judge-jury-and-executioner-the-unaccountable-algorithm [https://perma.cc/56VN-M3AT] ("Protected by trade secrecy, many algorithms remain impenetrable to outside observers.").

[47] Concededly, the Supreme Court has offered a First Amendment imprimatur for reuse of illegally obtained information in some contexts. See, e.g., Bartnicki v. Vopper, 532 U.S. 514, 517–18 (2001) (finding the First Amendment protects "speech that discloses the contents of an illegally intercepted communication"). However, that defense is conditioned on a "public interest" finding, id. at 540 (Breyer, J., concurring), and secret categorization or ranking of applicants should not qualify. See Frank Pasquale, Reforming the Law of Reputation, 47 Loy. U. Chi. L.J. 515, 529–30 (2015) (discussing the limits of *Bartnicki*).

Finally, certain inferences can become data that are extraordinarily suspect.[48] Consider, for instance, the rise of efforts to correlate persons' facial features and voices with illness, risk, or aptitude. Machine learning researchers have stirred controversy by claiming that our faces may reveal our sexual orientation and intelligence.[49] Using a database of prisoners' faces, some have even developed stereotypes of criminal features, reprising long-discredited physiognomy and phrenology.[50] A firm has claimed that it can deploy facial recognition to spot pedophiles and terrorists.[51] These inferences are deeply troubling. When such methods of pattern recognition are used to classify persons, they overstep a fundamental boundary between objective analysis and moral judgment. And when such moral judgments are made, persons categorized by the judgements deserve a chance to understand and contest them.

When a data set is not representative of the group it is used to classify, any results based on it should be clearly qualified. For example, a machine learning classifier may properly be said to succeed in classifying some percentage of faces *in its data set* in certain ways. But it should not be deployed as a potential classifier for all persons unless and until we have some sense of how the training set maps to the full set of persons it ostensibly classifies. As Dan McQuillan warns, machine learning often makes powerful predictions, "prompting comparisons with science. But rather than being universal and objective, it produces knowledge that is irrevocably entangled with specific computational mechanisms and the data used for training."[52] Both lawmakers and policymakers should hold users of such data sets responsible for making predictable errors based

[48] For a fuller account of the problem of troubling or inappropriate inferences, see Sandra Wachter & Brent Mittelstadt, A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI, 2019 Colum. Bus. L. Rev. 494, 499–505 (2019).

[49] See, e.g., Sam Levin, Face-Reading AI Will Be Able to Detect Your Politics and IQ, Professor Says, Guardian (Sept. 12, 2017), https://www.theguardian.com/technology/2017/sep/12/artificial-intelligence-face-recognition-michal-kosinski [https://perma.cc/X4HD-KNAK].

[50] Sam Biddle, Troubling Study Says Artificial Intelligence Can Predict Who Will Be Criminals Based on Facial Features, The Intercept (Nov. 18, 2016), https://theintercept.com/2016/11/18/troubling-study-says-artificial-intelligence-can-predict-who-will-be-criminals-based-on-facial-features/ [https://perma.cc/X3SN-QAEU]. It was later suggested that the sources of images used for the study may have been a key factor explaining its results. @davidjayharris, Twitter (Mar. 7, 2019), https://twitter.com/davidjayharris/status/1103636069180993537 [https://perma.cc/AKD5-TGPT].

[51] Matt McFarland, Terrorist or Pedophile? This Start-Up Says It Can Out Secrets by Analyzing Faces, Wash. Post (May 24, 2016), https://www.washingtonpost.com/news/innovations/wp/2016/05/24/terrorist-or-pedophile-this-start-up-says-it-can-out-secrets-by-analyzing-faces/ (on file with the *Columbia Law Review*).

[52] Dan McQuillan, People's Councils for Ethical Machine Learning, Soc. Media + Soc'y, Apr.–June 2018, at 1, 1, https://doi.org/10.1177/2056305118768303 [https://perma.cc/9CS7-4AAK].

on defective data sets, particularly if they fail to disclose the limitations of the data used.

II. COMPLEMENTARY TORT AND REGULATORY REGIMES

Tort law has evolved to handle the changing risks and affordances of new technologies.[53] However, judges alone cannot adequately respond to the new challenges posed by AI. Objective sources of information on best practices in data science are necessary as well. Expert agencies are particularly well positioned to analyze and articulate emerging industry standards, which should inform judicial determinations of standards of care. This Part describes emerging doctrinal and regulatory approaches that suggest data-driven duties for the developers of artificial intelligence. This type of data stewardship serves two purposes: ex ante, to ensure that the training data for machine learning adequately reflects the domain it governs or affects, and ex post, to detect anomalies and remedy them before they cause great harm.[54] Developing and maintaining these duties will be crucial to promoting just and humane advances in AI.

As Professors Dan Dobbs, Paul Hayden, and Ellen Bublick explain, "A tort is conduct that constitutes a legal wrong and causes harm for which courts will impose civil liability."[55] Negligence, vicarious liability, strict liability, and product liability regimes all may be relevant to future torts attributable to AI.[56] In the realm of negligence, the plaintiff generally must prove that the defendant caused the plaintiff's injury, owed a duty of care to the plaintiff, and breached that duty.[57] There are also diverse vicarious liability doctrines, each hinging on factors that include the degree of control an entity has over the direct cause of harm.[58] As

[53] Donald G. Gifford, Technological Triggers to Tort Revolutions: Steam Locomotives, Autonomous Vehicles, and Accident Compensation, 11 J. Tort L. 71, 143 (2018).

[54] Cf. Kristin Madison, Health Regulators as Data Stewards, 92 N.C. L. Rev. 1605, 1607–09 (2014) (arguing that regulators, as data stewards, bear a duty to serve as both an aggregator and editor of big health care data in order to ensure both the integrity of data collection and informed, continuous evaluation of regulation).

[55] Dan B. Dobbs, Paul T. Hayden & Ellen Bublick, Hornbook on Torts 3 (2d ed. 2016). This basic tort definition is consistent even in civil law countries around the world. See, e.g., Principles of European Tort Law: Text and Commentary tit. 1, art. 1:101 (Eur. Grp. on Tort Law 2005) ("Basic Norm (1) A person to whom damage to another is legally attributed is liable to compensate that damage."); Tort Law of the People's Republic of China, Ministry of Commerce of China (Dec. 26, 2009), http://english.mofcom.gov.cn/article/policyrelease/Businessregulations/201312/20131200432451.shtml [https://perma.cc/5LMK-YLDC] ("Those who infringe on civil rights and interests shall be subject to tort liability according to this Law.").

[56] For a useful typology of torts, see the table of contents of Dobbs et al., supra note 55, at xv–xxxi.

[57] Dan B. Dobbs, The Law of Torts 269 (2000).

[58] Harry Shulman, Fleming James Jr., Oscar S. Gray & Donald G. Gifford, Law of Torts: Cases and Materials 112–30 (5th ed. 2010).

services become more complex, one of the most promising developments in tort law is corporate liability for failure to maintain adequate safety standards.

For example, in one of the leading cases in medical corporate liability, *Thompson v. Nason Hospital*, the Pennsylvania Supreme Court did not allow the responsibility for a bad outcome to dissolve into a mist of contractual relationships among a hospital, its staff, doctors, and the manufacturers of devices that its doctors and staff used.[59] Rather, the *Thompson* court articulated a general duty of a hospital "to ensure the patient's safety and well-being while at the hospital."[60] The court went on to articulate four nonexhaustive dimensions of this general duty to protect safety and well-being:

> (1) a duty to use reasonable care in the maintenance of safe and adequate facilities and equipment; (2) a duty to select and retain only competent physicians; (3) a duty to oversee all persons who practice medicine within its walls as to patient care; and (4) a duty to formulate, adopt and enforce adequate rules and policies to ensure quality care for the patients.[61]

This standard of corporate negligence has much to offer outside of the healthcare setting. One classic theoretical foundation of health law as a distinctive field is the great difference between ordinary consumer markets, on the one hand, and the healthcare field, where information asymmetries and power differentials routinely arise between patients and healthcare providers, on the other.[62] The rise of software and cyber-physical

[59] 591 A.2d 703, 709 (Pa. 1991).

[60] Id. at 707.

[61] Id. (citations omitted).

[62] As Donald Cohodes argues, medical care can be differentiated from "most other products" in six general ways:

> 1. *Demand for health.* Medical care services are not purchased from any desire for such services in themselves . . . [but instead are] derived from the "demand" for good health.
>
> 2. *Medical care and health.* Medical care is only one determinant of health status, and for most people at most times it is not even a very important determinant. . . .
>
> 3. *Risk.* The need for medical care is unpredictable, requiring expenditures that are irregular and of uncertain magnitude.
>
> 4. *Immediacy.* The need for medical care is often immediate, allowing little time for shopping around and seeking advice or alternatives.
>
> 5. *Lack of Information.* Consumers are usually ignorant of their medical care needs. They cannot possibly obtain the knowledge and training to diagnose their own medical care needs . . . .
>
> 6. *Uncertainty.* Physicians, though highly trained and better able to diagnose needs and prescribed treatment, also are often uncertain about the appropriate services to provide.

Donald R. Cohodes, Where You Stand Depends on Where You Sit: Musings on the Regulation/Competition Dialogue, 7 J. Health Pol. Pol'y & L. 54, 56 (1982).

systems portends a similar increase in complexity, power differentials, and information asymmetry reminiscent of the highly scientific and professionalized medical milieu.[63] Doctrines and approaches developed in the medical setting have already been proposed for other aspects of data governance. For example, health privacy law can serve as a model for the regulation of other data.[64] Jack Balkin and Jonathan Zittrain have proposed that a law of fiduciary duties, itself heavily reliant on the model of doctors' duties to patients, should bind large technology firms with respect to their treatment of data collected from users.[65]

Thompson has been cited many times, and its factors helpfully articulate theories of liability.[66] An elaboration of the corporate negligence standard in a complex environment can illuminate the roles and responsibilities of the developers of artificial intelligence. For example, the first duty (to use reasonable care in the maintenance of safe facilities and equipment) suggests a similar obligation to exercise due care in the selection of sources of data. *Thompson* also reflects in law the conclusions of a larger quality-improvement movement: that it is less important to find particular persons to blame in the case of accidents, than to identify malfunctioning sociotechnical systems of human–computer interaction.[67]

The third *Thompson* factor, regarding adequate supervision, also raises important questions in the context of automation developed in corporate labs and its testing outside of controlled settings. Surveillance techniques are widespread and well-developed.[68] Such technology could

[63] On the rise of software in ordinary products, see Paul Ohm & Blake Reid, Regulating Software When Everything Has Software, 84 Geo. Wash. L. Rev. 1672, 1676–79 (2016); see also James Grimmelmann, Note, Regulation by Software, 114 Yale L.J. 1719, 1723–24 (2005) (giving "four patterns [that] provide a general methodology for assessing the use of software in a given regulatory context").

[64] Frank Pasquale, The Black Box Society: The Secret Algorithms that Control Money and Information 150–51 (2015) (discussing HIPAA standards for consent, security, and accounting of disclosures of health data as a model for other forms of data).

[65] Jack M. Balkin, Information Fiduciaries and the First Amendment, 49 U.C. Davis L. Rev. 1183, 1221–25 (2016). But see Lina Khan & David Pozen, A Skeptical View of Information Fiduciaries, 133 Harv. L. Rev. (forthcoming 2019) (manuscript at 6–8), https://ssrn.com/abstract=3341661 (on file with the *Columbia Law Review*) (arguing that creating new fiduciary duties based on information custody is fundamentally incompatible with existing corporate law of fiduciary duties and therefore impossible to implement in the form proposed by Balkin and Zittrain).

[66] As of March 15, 2019, Thompson v. Nason Hospital, 591 A.2d 703 (Pa. 1991), has been cited in 198 cases and 273 secondary sources on Westlaw Edge.

[67] Lucian L. Leape, Error in Medicine, 272 JAMA 1851, 1853 (1994) (describing the importance of system-level analysis in attribution of blame and prevention of future harms).

[68] See, e.g., Karen E.C. Levy, The Contexts of Control: Information, Power, and Truck-Driving Work, 31 Info. Soc'y 160, 160, 164 (2015) (describing how trucking firms have extensively deployed telematics to monitor truck drivers with regard to performance and timekeeping); Steve Kolowich, Behind the Webcam's Watchful Eye, Online Proctoring Takes Hold, Chron. Higher Educ. (Apr. 15, 2013), https://www.chronicle.com/article/

help reduce bias in data collection and promote vigilance among those tasked with overseeing the deployment of AI in sensitive settings. On the other hand, privacy activists may raise concerns if the common law of tort promotes excessive surveillance of workers.[69] Once again, the health care industry has been at the forefront, developing balanced frameworks for the inclusion of surveillance technology in workplaces in which human life is routinely at risk.[70]

III. Regulatory Standards for Data Use and Reporting

If a large proportion of cases involving AI went to trial, reported opinions would serve as a prominent source of guidance for AI vendors and users concerned about safety and effectiveness. However, we can expect that here, as with data security, the prevalence of settlements of disputes will frustrate such evolutionary clarification of duties.[71] In this vacuum, regulators should play a vital role in setting (or at least informing)

Behind-the-Webcams-Watchful/138505 [https://perma.cc/CQ5T-2C74] (describing online proctors that watch students through a webcam to detect cheating); Natasha Singer, Online Test-Takers Feel Anti-Cheating Software's Uneasy Glare, N.Y. Times (Apr. 5, 2015), https://www.nytimes.com/2015/04/06/technology/online-test-takers-feel-anti-cheating-softwares-uneasy-glare.html (on file with the *Columbia Law Review*) (describing software developed to detect cheating during online and computer exam taking).

[69] See, e.g., Lewis Maltby, Can They Do That?: Retaking Our Fundamental Rights in the Workplace 16–17 (2009) (describing an example of intrusive surveillance of workers); Ifeoma Ajunwa, Kate Crawford & Jason Schultz, Limitless Worker Surveillance, 105 Calif. L. Rev. 735, 735–36, 772–73 (2017) (describing the trend of increased worker surveillance and exploring possible remedies to protect worker privacy).

[70] See generally Clara Berridge, Jodi Halpern & Karen Levy, Cameras on Beds: The Ethics of Surveillance in Nursing Home Rooms, 10 AJOB Empirical Bioethics 55 (2019) (examining survey data on the use of "family-provided cameras" in nursing homes and their legal and ethical implications); Karen Levy, Lauren Kilgour & Clara Berridge, Regulating Privacy in Public/Private Space: The Case of Nursing Home Monitoring Laws, 26 Elder L.J. 323, 326–27 (2019) (comparing "state laws and regulations governing resident-room cameras in nursing homes . . . focus[ing] on how such rules approach and balance the privacy concerns of the multiple relations involved in such contexts, and how legal protections do—and do not—address relationship-specific interests").

[71] See William McGeveran, The Duty of Data Security, 103 Minn. L. Rev. 1135, 1144 (2019) ("There are numerous lawsuits about data security, which raise claims under tort, contract, or consumer protection law, among other theories. Courts considering these cases offer hardly any insight into the *content* of the duty of data security, however, because they almost never reach the merits." (footnote omitted)); cf. Owen M. Fiss, Against Settlement, 93 Yale L.J. 1073, 1075, 1078–85 (1984) (complaining of the problems caused by this avoidance). Instead, in the data security context, the Federal Trade Commission has taken the lead. See Woodrow Hartzog and Daniel J. Solove, The FTC and the New Common Law of Privacy, 114 Colum. L. Rev. 583, 585–86 (2014) ("Despite over fifteen years of FTC enforcement, there are hardly any judicial decisions to show for it. The cases have nearly all resulted in settlement agreements. . . . Thus, in practice, FTC privacy jurisprudence has become the broadest and most influential regulating force on information privacy in the United States . . . .").

standards.[72] Though the current Congress is unlikely to establish a new agency, existing statutory authorities already grant extant agencies the power to gather, analyze, and disseminate data that would aid courts' assessments of the proper standard of care in disputes related to AI-informed and AI-performed services.[73] Some of these agencies have also established standards that have informed tort cases in data-related fields, such as privacy law.[74]

A. *Ensuring the Integrity of Inputs*

One purpose of the Health Insurance Portability and Accountability Act's (HIPAA) security requirements is to protect data from hackers or other corrupting influences.[75] A logical extension of this duty is for agencies to set standards for AI vendors and users to verify the quality and accuracy of the data they use.[76] These standards may start at an elementary level. For example, HIPAA best practices dictate that a covered entity both record any source of data it receives and record its transfer of data to other covered entities or business associates.[77] Those recipients of data must in turn do the same.[78] This creates a set of links that makes it easier to trace and then minimize the impact of inaccurate, unrepresentative,

[72] For a general account of the government role in promoting standardized data, see generally Michal S. Gal & Daniel L. Rubinfeld, Data Standardization, 94 N.Y.U. L. Rev. (forthcoming 2019), https://ssrn.com/abstract=3326377 (on file with the *Columbia Law Review*).

[73] See Andrew F. Popper, Gwendolyn M. McKee, Anthony E. Varona, Philip J. Harter, Mark C. Niles & Frank Pasquale, Administrative Law: A Contemporary Approach 1067–134 (3d ed. 2016) (describing the power, and the limits of such power, of U.S. agencies to demand information).

[74] See infra sections III.A–.B.

[75] See Frank Pasquale, Redescribing Health Privacy: The Importance of Information Policy, 14 Hous. J. Health L. & Pol'y 95, 105–09 (2014) (describing the range of security measures prescribed by HIPAA).

[76] See, e.g., Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz & Oscar Schwartz, AI Now Report 2018, at 4–7 (2018), https://ainowinstitute.org/AI_Now_2018_Report.pdf [https://perma.cc/4J3T-TCTR] (discussing the importance of sectoral regulation); cf. Frank Pasquale, Private Certifiers and Deputies in American Health Care, 92 N.C. L. Rev. 1661, 1668–69, 1671–73, 1692 (2014) (describing a broad array of public and private actors that have cooperated in highly technical areas to promote data quality and interoperability in the health care industry).

[77] Bill Becker, HIPAA Compliance Best Practices: Questions and Answers to Improve Security and Avoid Penalties, HIPAA J. (May 16, 2017), https://www.hipaajournal.com/hipaa-compliance-best-practices-8809/ [https://perma.cc/GZZ5-HM4L]; Office for Civil Rights, How Are Covered Entities Expected to Determine What Is the Minimum Necessary Information that Can Be Used, Disclosed, or Requested for a Particular Purpose?, HHS: Health Info. Privacy (Dec. 19, 2002), https://www.hhs.gov/hipaa/for-professionals/faq/207/how-are-covered-entities-to-determine-what-is-minimum-necessary/index.html [https://perma.cc/R788-UZRW] (last updated Mar. 14, 2006).

[78] See Becker, supra note 77 (discussing best practices for improving data security); Office for Civil Rights, supra note 77 (setting out requirements for minimum data sharing).

or otherwise compromised data.[79] Similar standards should inform the stewardship of data used for machine learning and AI. Federal standards for data protection may, in turn, become part of the standard of care for torts like breach of medical confidentiality.[80]

For a concrete example of why such practices matter, consider how voice recognition software may be more or less accurate with respect to persons with different voices or accents.[81] As of 2020, databases may have a certain level of inclusiveness;[82] by 2025, this is likely to have improved markedly.[83] An AI vendor using the 2020 database in 2025 for mission-critical applications may rightly be faulted for failing to update in light of new knowledge about the limitations of the database. But we would not even know where to look for such a problem if the source of the firm's data was not recorded adequately.[84]

[79] See Woodrow Hartzog, Chain Link Confidentiality, 46 Ga. L. Rev. 657, 677 (2012) ("The HIPAA Privacy Rules provide that, although only covered entities such as healthcare providers are bound to confidentiality, these entities may not disclose information to their business associates without executing a written contract that places the business associate under the same confidentiality requirements as the healthcare providers."). These protections have been strengthened even further by the Health Information Technology for Economic and Clinical Health Act (HITECH) (and the HIPAA Omnibus Rule of 2013), which impose statutory and regulatory duties on business associates and even their downstream contractors. See Frank Pasquale & Tara Adams Ragone, Protecting Health Privacy in an Era of Big Data Processing and Cloud Computing, 17 Stan. Tech. L. Rev. 595, 609–15 (2014) (describing these duties).

[80] See Barry R. Furrow, Thomas L. Greaney, Sandra H. Johnson, Timothy Stoltzfus Jost, Robert L. Schwartz, Brietta R. Clark, Erin C. Fuse Brown, Robert Gatter, Jaime S. King & Elizabeth Pendo, Health Law: Cases, Materials and Problems 201 (8th ed. 2018) (noting that courts have held that "despite the absence of a private right of action under HIPAA, it can inform the applicable standard of care in common law tort cases"); see also Bonney v. Stephens Mem'l Hosp., 17 A.3d 123, 128 (Me. 2011) ("HIPAA standards, like state laws and professional codes of conduct, may be admissible to establish the standard of care associated with a state tort claim . . . ."); Acosta v. Byrum, 638 S.E.2d 246, 253 (N.C. Ct. App. 2006) (describing HIPAA "providing evidence of the duty of care owed . . . with regards to the privacy of plaintiff's medical records"). But see Young v. Carran, 289 S.W.3d 586 (Ky. Ct. App. 2008) (declining to adopt a negligence per se standard); Sheldon v. Kettering Health Network, 40 N.E.3d 661, 664 (Ohio Ct. App. 2015) (same).

[81] See Sonia Paul, Voice Is the Next Big Platform, Unless You Have an Accent, WIRED (Mar. 20, 2017), https://www.wired.com/2017/03/voice-is-the-next-big-platform-unless-you-have-an-accent/ [https://perma.cc/78TL-PSC9] (reporting on the difficulties associated with creating software that recognizes different accents).

[82] See id. (reporting tech companies' efforts to improve the inclusiveness of their accent data); Kyle Wiggers, These Companies Are Shrinking the Voice Recognition 'Accent Gap,' Venture Beat (Aug. 11, 2018), https://venturebeat.com/2018/08/11/using-ai-and-big-data-to-address-the-accent-gap-in-voice-recognition-systems/ [https://perma.cc/F96Z-9FH4] (same).

[83] See Paul, supra note 81; Wiggers, supra note 82.

[84] This is not a mere hypothetical; I recently had to take to Twitter to learn where the voices for a Google Assistant feature (Duplex) came from. The source was not clearly labeled on the corporate website trumpeting the feature.

Such standards will be resisted. AI vendors will likely push for another approach, simply disclosing potential problems with their data in advance in disclaimers.[85] Perhaps it is the responsibility of the person using the AI, rather than the vendor of AI, to correct for error-prone datasets. However, courts may also find ample precedent for holding vendors responsible. For example, in lawsuits over food poisoning, consumers' "reasonable expectation" of purity and appropriateness of ingredients has been recognized.[86]

Some AI-driven devices may also need to be subjected to the certification and testing now applied (albeit minimally) to electronic health records.[87] Thanks to the HITECH Act of 2009, the Department of Health and Human Services must assure that EHRs meet basic functionality requirements.[88] Failures of EHR vendors to comply with federal health standards have already led to litigation.[89] Given the False Claims Act's

[85] For entertaining examples of the rhetoric one can expect, see Chris Jay Hoofnagle, Denialists' Deck of Cards: An Illustrated Taxonomy of Rhetoric Used to Frustrate Consumer Protection Efforts (Feb. 9, 2007) (unpublished manuscript), https://ssrn.com/abstract=962462 (on file with the *Columbia Law Review*) (illustrating "a taxonomy of arguments used in denialism" by using "a deck of playing cards to make it more interesting and to emphasize that denialists are engaged in a predictable game to 'do little and delay.'").

[86] See Gail Kachadurian McCallion, Note, From the Source to the Mouth: What Can You Reasonably Expect to Find in Your Food?, 5 Fordham Envtl. L.J. 189, 212 (1993) ("The reasonable expectation test asserts that regardless of whether a substance in a food product is natural to an ingredient, liability will lie for injuries caused by the substance where the consumer of the product would not have reasonably expected to find the substance in the product."); see also Richard E. Kaye, Foreign Substance in Food or Beverage, 124 Am. Jur. Proof Facts 3d 91, § 2 (2018) (citing Restatement (Third) of Torts: Prod. Liab. § 7 (Am. Law Inst. 1998)).

[87] See, e.g., ONC—Authorized Testing Laboratories (ONC-ATLs), HealthIT.gov (Sept. 24, 2018), https://www.healthit.gov/node/95011 [https://perma.cc/63XV-7A97] (listing the five Authorized Testing Laboratories "accredited by NVLAP and authorized by ONC to test Health IT Modules under the ONC Health IT Certification Program").

[88] Health Information Technology Standards, 45 C.F.R. §§ 170.202–170.210 (2019) (providing a detailed set of standards for the use and storage of electronic health information including, for example, encryption and hashing algorithm requirements).

[89] See, e.g., Complaint at 1–4, United States ex rel. Delaney v. eClinicalWorks, No. 2:15-CV-00095-WKS (D. Vt. May 1, 2015) (alleging that the defendant failed to comply with federal requirements and that it misrepresented information and failed to disclose flaws in its EHR system in violation of the False Claims Act). The defendant later settled the claim for $155 million. See Press Release, Dep't of Justice, Electronic Health Records Vendor to Pay 155 Million to Settle False Claims Act Allegations (May 31, 2017), https://www.justice.gov/opa/pr/electronic-health-records-vendor-pay-155-million-settle-false-claims-act-allegations [https://perma.cc/59B3-6M5E]; see also Jessica Davis, eClinicalWorks Sued for Nearly $1 Billion for Inaccurate Medical Records, Healthcare IT News (Nov. 17, 2017), https://www.healthcareitnews.com/news/eclinicalworks-sued-nearly-1-billion-inaccurate-medical-records [https://perma.cc/3FFF-NNQV] ("EHR vendor eClinicalWorks has been hit with a class-action lawsuit that alleges . . . that millions of patients have compromised patient records, as eClinicalWorks' software didn't meet meaningful use and certification requirements laid out by the Office of the National Coordinator."); Heather Landi, $1 Billion Class Action Lawsuit Filed Against eClinicalWorks, Healthcare Innovation (Nov. 20, 2017),

(FCA) role in assuring that healthcare providers are treating patients with valid and effective forms of care, this form of liability should be a bellwether specifically for AI vendors contracting with governmental authorities. Consumer protection authorities should also take note.

B. *Ensuring the Transparency of Outputs*

Health regulators have long considered data stewardship a critical role under their statutory mandate.[90] When the federal government began funding EHRs in earnest in 2011, it not only demanded certain basic recordkeeping but also set providers on an ambitious path toward "meaningful use" of information technology—including potentially AI-driven tools like clinical decision support.[91] In 2015, Congress promoted interoperability in the Medicare Access and CHIP Reauthorization Act (MACRA).[92] This drive for interoperability continues to this day, as the Office for the National Coordinator of Health Information Technology and the Centers for Medicare and Medicaid Services have recently announced rulemakings designed to help promote data liquidity.[93]

One key rationale for interoperability is supporting the massive disclosure and reporting requirements mandated pursuant to healthcare finance reforms (covering Advanced Payment Models (APMs) such as Accountable Care Organizations (ACOs), as well as readmissions penalties and bundled payments).[94] It may be very difficult for networks like

https://www.hcinnovationgroup.com/clinical-it/news/13029475/1-billion-class-action-lawsuit-filed-against-eclinicalworks [https://perma.cc/EKR7-KPKJ] ("The class action lawsuit alleges ECW falsely represented to its certifying bodies that its software complied with the requirements for certification and the payment of incentives under the MU program, and therefore, caused its users to falsely attest to using a certified EHR technology.").

[90] See generally Madison, supra note 54, at 1607–28 (discussing ways the federal government has taken on "the responsibility for protecting the integrity and confidentiality of data" in the health care sector).

[91] Frank Pasquale, Grand Bargains for Big Data: The Emerging Law of Health Information, 72 Md. L. Rev. 682, 710–11 (2013) (explaining how the law promotes patient health).

[92] Pub. L. No. 114-10, § 106(b), 129 Stat. 87, 138–40 (2015) (codified at 42 U.S.C. § 1395w-4 (2018)) ("The term 'interoperability' means the ability of two or more health information systems or components to exchange clinical and other information . . . to provide access to longitudinal information for health care providers in order to facilitate coordinated care and improved patient outcomes.").

[93] Interoperability and Patient Access for Medicare Advantage Organization and Medicaid Managed Care Plans, State Medicaid Agencies, CHIP Agencies and CHIP Managed Care Entities, Issuers of Qualified Health Plans in the Federally-Facilitated Exchanges and Health Care Providers, 84 Fed. Reg. 7610 (proposed Mar. 4, 2019); 21st Century Cares Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program, 84 Fed. Reg. 7424 (proposed Mar. 4, 2019).

[94] See, e.g., 2019 Program Requirements Medicare, Ctrs. for Medicare & Medicaid Servs., https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/2019ProgramRequirementsMedicare.html [https://perma.cc/GB7L-PK34] (last modified May 8, 2019) (describing reporting requirements for Medicare programs to comply with Promoting Interoperability measures); MIPS Overview, Quality Payment Program, Ctrs. for

ACOs to accurately report on quality standards without a common infrastructure of EHRs that can aggregate data on key performance indicators and benchmarks.[95] A common indicator of nosocomial infection, for instance, may be critical to ensuring the integrity of performance assessment.

AI applications are already playing a role in promoting health-related interventions and should be subject to similar performance assessments. For example, as Natasha Singer has reported, Facebook has deployed an algorithm to flag users that may be so suicidal that police should be called by Facebook employees to intervene.[96] Mason Marks has documented numerous other examples of "social suicide prediction" programs, which use machine learning to generate risk scores for individuals.[97] There are long-term risks to privacy and autonomy that such scores could create—for example, if unregulated and shared beyond their source, they may affect the marketing a person experiences, or even job or insurance opportunities.[98]

They also raise important concerns about immediate risks to safety caused by false positives. What are the stigmatic concerns raised by being falsely accused of extreme suicidality, or of a suicide attempt? What do first responders think of the interventions they have been prompted to carry out? Ensuring that there are standard ways of reporting positive and negative interventions here could help policymakers better determine which AI to fund in this critical area. It could also nip in the bud problematic interventions, like the Samaritans' Radar App, which shut

Medicare & Medicaid Servs., https://qpp.cms.gov/mips/overview [https://perma.cc/BWX7-YNEM] (last visited June 28, 2019) (describing the Merit-based Incentive Payment System (MIPS) and four areas of reporting: "Quality, Improvement Activities, Promoting Interoperability (formerly Advancing Care Information), and Cost"); Promoting Interoperability (PI), Ctrs. for Medicare & Medicaid Servs., https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html?redirect=/EHRincentiveprograms [https://perma.cc/CH85-HQ5Q] (last modified Aug. 14, 2019) (describing CMS' Promoting Interoperability program).

[95] On the role of such indicators and benchmarks in AI-driven medical practice, see Frank Pasquale, Professional Judgment in an Era of Artificial Intelligence and Machine Learning, 46 boundary 2, at 73, 85 (2019) (exploring the role that metrics play in ACO performance assessments and compensation under the Affordable Care Act).

[96] Natasha Singer, In Screening for Suicide Risk, Facebook Takes on Tricky Public Health Role, N.Y. Times (Dec. 31, 2018), https://www.nytimes.com/2018/12/31/technology/facebook-suicide-screening-algorithm.html (on file with the *Columbia Law Review*).

[97] See generally Mason Marks, Artificial Intelligence Based Suicide Prediction, Yale J. Health Pol'y L. & Ethics (forthcoming), https://ssrn.com/abstract=3324874 (on file with the *Columbia Law Review*) (discussing the contours and unforeseen consequences of programs initiated by companies such Facebook, Crisis Text Line, and Operation Zero that "collect [consumers'] digital traces and analyze them with AI to infer [consumers'] health information").

[98] Danielle Keats Citron & Frank Pasquale, The Scored Society: Due Process for Automated Predictions, 89 Wash. L. Rev. 1, 4 (2014) (describing spread of scoring technologies).

down its simple program for automated detection of suicidality after public complaints.[99]

CONCLUSION

Futurists envision AI programs that effectively act of their own accord, without direction or control by their developers (or any other person). Such entities could be quite dangerous.[100] However, advocates for such AI believe that law should effectively step out of the way of its development. How, the question goes, can the creators or owners of such general-purpose technology anticipate all the potential legal problems their AI might generate or encounter? No one wants to hold Microsoft responsible for ransom notes written with MSWord—it is a blank slate. Nor are parents responsible for the crimes of their children—they are independent entities.

Leading developers of AI, at present, benefit from both the "blank slate" and "independent entity" intuitions of nonresponsibility for their creations. But neither should immunize such firms, given a decade of research on algorithmic accountability. As Jack Balkin has observed, we all now know that algorithms can "(a) construct identity and reputation through (b) classification and risk assessment, creating the opportunity for (c) discrimination, normalization, and manipulation, without (d) adequate transparency, accountability, monitoring, or due process."[101] Moreover, we are well aware of their ability to malfunction, dating back at least to the Therac-25 debacle of the 1980s.[102] These factors all counsel in favor of discouraging the development of any AI whose actions are not directly attributable to a person or persons that can be held responsible for them.[103]

[99] Jamie Orme, Samaritans Pulls 'Suicide Watch' Radar App over Privacy Concerns, Guardian (Nov. 7, 2014), https://www.theguardian.com/society/2014/nov/07/samaritans-radar-app-suicide-watch-privacy-twitter-users [https://perma.cc/342U-99KP].

[100] See, e.g., Lynn M. Lopucki, Algorithmic Entities, 95 Wash. U. L. Rev. 887, 951 (2018) ("[Algorithmic entities] constitute a threat to humanity because the only limits on their conduct are the limits the least restrictive human creator imposes.").

[101] Jack M. Balkin, 2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data, 78 Ohio St. L.J. 1217, 1239 (2017). Algorithmic information processing is in effect the "brain" of robotics and AI agents. See generally Domingos, supra note 41, at 93–119.

[102] See generally Edmond W. Israelski & William H. Muto, Human Factors Risk Management as a Way to Improve Medical Device Safety: A Case Study of the Therac 25 Radiation Therapy System, 30 Joint Commission J. Quality & Safety 689 (2004); Nancy G. Leveson & Clark S. Turner, An Investigation of the Therac-25 Accidents, Computer, July 1993, at 18, 18 (presenting an accident investigation of overdoses caused by the Therac-25 radiation therapy machine).

[103] Frank Pasquale, Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society, 78 Ohio St. L.J. 1243, 1252–55 (2017) (arguing that very high levels of autonomy should be illegal if any harm is foreseeable, given the problems of attribution they can give rise to).

However appealing dreams of artificial general intelligence may be, the dominant version of AI now prevalent in commerce and government is only a few steps removed from algorithmic systems we are all now familiar with. For example, "AI hiring" based on voice parsing is not a substitute for a Director of Human Resources.[104] Nor is it an all-purpose assessment of character. Rather, it is a method of translating data (a voice) into an output (an assessment of likely success at a job) based on computational analysis of how past employees with similar voices have fared at the job. True, the concept of "similarity" here may have far more dimensions than a simple linear relationship; contemporary machine learning is premised on advances in computational power that not only allow various, granular hypotheses to be tested, but also combine potentially relevant variables in myriad ways.[105] However, the collection, analysis, and use of data is foundational to the process, and presents several opportunities for imposing duties on AI developers, given possibly inaccurate or inappropriate data.

Advocates for legal technology (including legaltech, regtech, and fintech) have promoted a "duty of technological competence" for lawyers.[106] In many cases, an attorney cannot properly serve a client without knowing how to use certain databases or search engines. Nor can a lawyer competently advise a modern business on a topic like document retention without a clear sense of how computers store data. Rules of

[104] Stephen Buranyi, How to Persuade a Robot that You Should Get the Job, Guardian (Mar. 3, 2018), https://www.theguardian.com/technology/2018/mar/04/robots-screen-candidates-for-jobs-artificial-intelligence [https://perma.cc/5GCW-JN5E]. For further descriptions of such analytics, see Matthew T. Bodie, Miriam A. Cherry, Marcia L. McCormick & Jintong Tang, The Law and Policy of People Analytics, 88 U. Colo. L. Rev. 961, 963, 1032–38 (2017).

[105] See The Royal Soc'y, Machine Learning: The Power and Promise of Computers that Learn by Example 19–20 (2017), https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf [https://perma.cc/YUE9-87MZ].

[106] See, e.g., Model Rules of Prof'l Conduct 1.1 cmt. 8 (Am. Bar Ass'n 2018) (including the duty to "maintain the requisite knowledge and skill . . . including [keeping up-to-date on] the benefits and risks associated with relevant technology"); see also Anthony E. Davis & Steven M. Puiszis, An Update on Lawyers' Duty of Technological Competence: Part 1, N.Y. L.J. (Mar. 1, 2019), https://www.law.com/newyorklawjournal/2019/03/01/an-update-on-lawyers-duty-of-technological-competence-part-1/ [https://perma.cc/3XU5-6P4P] [hereinafter Davis & Puiszis, Update Part 1] (arguing that it is necessary for today's lawyers to maintain data security and become familiar with the technology used to run a law firm and practice law); Anthony E. Davis & Steven M. Puiszis, An Update on Lawyers' Duty of Technological Competence: Part 2, N.Y. L.J. (May 3, 2019), https://www.law.com/newyorklawjournal/2019/05/03/an-update-on-lawyers-duty-of-technological-competence-part-2/ [https://perma.cc/P7Z5-BWX4] [hereinafter Davis & Puiszis, Update Part 2] (using social media, electronic discovery, client technology, and technology to present information in court).

professional responsibility, as well as tort doctrines of legal malpractice,[107] enforce a duty of technological competence on many attorneys.[108]

In numerous fields, there is a parallel duty for technology providers to have some basic understanding of the law as they serve their clients. A video hosting service in the United States, for example, needs to understand the fundamentals of copyright law.[109] Firms developing electronic health record software unaware of the requirements of HIPAA[110] (and many other laws governing health privacy) cannot serve their clients well. In these cases, and many others, the onus is not simply on the buyer of the technology to vet what it is buying or leasing. Rather, principles of secondary liability effectively impose what might be called a duty of legal competence—of a basic understanding of what law requires—on technologists.[111] Some popular understandings of artificial intelligence pose a threat to the duty of legal competence by mystifying the bases of decisions. However, law and policy can require basic safeguards be taken in its development, can standardize public reporting on its effectiveness and safety, and can impose liability on the developers of unsafe, biased, or otherwise defective AI.

The promise of AI law and policy is to ensure that the owners and developers of algorithms are more accountable to the public.[112] Without imposing legal duties on the developers of AI, there is little chance of ensuring accountable technological development in this field. By focusing on data, the fundamental input for AI, both judges and policymakers can channel the development of AI to respect, rather than evade, core legal values.

[107] See, e.g., James v. Nat'l Fin. LLC, No. 8931-VCL, 2014 WL 6845560, at *12 (Del. Ch. Dec. 5, 2014) (citing Del. Rules of Prof'l Conduct 1.1 cmt. 8).

[108] See Model Rules of Prof'l Conduct r. 1.1 cmt. 8; Katherine Medianik, Note, Artificially Intelligent Lawyers: Updating the Model Rules of Professional Conduct in Accordance with the New Technological Era, 39 Cardozo L. Rev. 1497, 1512, 1514–16 (2018); Davis & Puiszis, Update Part 1, supra note 106; Davis & Puiszis, Update Part 2, supra note 106.

[109] For an example of such copyright law, see Digital Millennium Copyright Act of 1998, 17 U.S.C. §§ 1202–1332 (2012).

[110] Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (codified in scattered sections of 26, 29, and 42 U.S.C.).

[111] In the case of HIPAA, the secondary liability would be imposed on the vendor via a business associate agreement. See Pasquale & Adams Ragone, supra note 79, at 609–15.

[112] See Robyn Caplan, Joan Donovan, Lauren Hanson, & Jeanna Mathews, Algorithmic Accountability: A Primer 10 (2018), https://datasociety.net/output/algorithmic-accountability-a-primer/ [https://perma.cc/UTW2-62M9] ("Algorithmic accountability ultimately refers to the assignment of responsibility for how an algorithm is created and its impact on society; if harm occurs, accountable systems include a mechanism for redress."). Edward Rubin has defined accountability as "the ability of one actor to demand an explanation or justification of another actor for its actions, and to reward or punish the second actor on the basis of its performance or its explanation." Edward Rubin, The Myth of Accountability and the Anti-Administrative Impulse, 103 Mich. L. Rev. 2073, 2073 (2005).

AI SYSTEMS AS STATE ACTORS

Kate Crawford & Jason Schultz***

Many legal scholars have explored how courts can apply legal doctrines, such as procedural due process and equal protection, directly to government actors when those actors deploy artificial intelligence (AI) systems. But very little attention has been given to how courts should hold private vendors of these technologies accountable when the government uses their AI tools in ways that violate the law. This is a concerning gap, given that governments are turning to third-party vendors with increasing frequency to provide the algorithmic architectures for public services, including welfare benefits and criminal risk assessments. As such, when challenged, many state governments have disclaimed any knowledge or ability to understand, explain, or remedy problems created by AI systems that they have procured from third parties. The general position has been "we cannot be responsible for something we don't understand." This means that algorithmic systems are contributing to the process of government decisionmaking without any mechanisms of accountability or liability. They fall within an accountability gap.

In response, we argue that courts should adopt a version of the state action doctrine to apply to vendors who supply AI systems for government decisionmaking. Analyzing the state action doctrine's public function, compulsion, and joint participation tests, we argue that—much like other private actors who perform traditional core government functions at the behest of the state—developers of AI systems that directly influence government decisions should be found to be state actors for purposes of constitutional liability. This is a necessary step, we suggest, to bridge the current AI accountability gap.

INTRODUCTION

Advocates and experts are increasingly concerned about the rapid introduction of artificial intelligence (AI) systems in government services, from facial recognition and autonomous weapons to criminal risk

* Distinguished Research Professor, New York University; Principal Researcher, Microsoft Research; Co-Founder, AI Now Institute at New York University.

** Professor of Clinical Law, New York University; Law & Policy Area Lead, AI Now Institute at New York University. The authors wish to thank the participants in the *Columbia Law Review*'s 2019 Symposium "Common Law for the Age of AI" for their valuable feedback, and especially Frank Pasquale and Eric Talley for their insightful comments. They also wish to thank Eli Hadley for her excellent research help.

assessments and public benefits administration.[1] Every month, more algorithmic and predictive technologies are being applied in domains such as healthcare, education, criminal justice, and beyond.[2] A range of "advocates, academics, and policymakers have raised serious concerns over the use of such systems, which are often deployed without adequate assessment, safeguards, [or] oversight."[3] This is due, in part, to the fact that government agencies commonly outsource the development—and sometimes the implementation—of these systems to third-party vendors.[4] This outsourcing often leaves public officials and employees without any real understanding of those systems' inner workings or, more importantly, the variety of risks they might pose. Such risks range from discrimination and disparate treatment to lack of due process, discontinuance of essential services, and harmful misrepresentations.[5]

These risks are neither hypothetical nor intangible. Today, AI systems help governments decide everything from whom to release on bail,[6]

[1] See Litigating Algorithms, AI Now Inst., (Sept. 24, 2018), https://ainowinstitute.org/announcements/litigating-algorithms.html [https://perma.cc/683L-NSBJ] [hereinafter Litigating Algorithms Announcement]; infra section I.A. The term "artificial intelligence" has taken on many meanings, especially in conversations about law and policy. For this Essay, we will use it as a broad umbrella term, covering any computational system that utilizes machine learning, including deep learning and reinforcement learning; neural networks and algorithmic decisionmaking; and other similar techniques to generate predictions, classifications, or determinations about individuals or groups. We choose this definition in part because, while some of the systems we discuss may not actively incorporate the most modern AI techniques, they are designed with the same objectives in mind and aim to usher in AI capabilities as soon as they are feasible or available.

[2] See infra section I.A.

[3] Litigating Algorithms Announcement, supra note 1.

[4] See, e.g., AI Now Inst., Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems 7 (2018), https://ainowinstitute.org/litigatingalgorithms.pdf [https://perma.cc/FSG5-JBHT] [hereinafter Litigating Algorithms] ("Government agencies adopting these systems commonly enter into contracts with third-party vendors that handle everything.").

[5] For a survey of these risks and concerns, see generally Solon Barocas & Andrew D. Selbst, Big Data's Disparate Impact, 104 Calif. L. Rev. 671 (2016) (using the lens of antidiscrimination law to explore bias arising from data mining); Danielle Keats Citron & Frank Pasquale, The Scored Society: Due Process for Automated Predictions, 89 Wash. L. Rev. 1 (2014) (warning that additional procedural safeguards are necessary for automated prediction systems); Danielle Keats Citron, Technological Due Process, 85 Wash. U. L. Rev. 1249 (2008) (proposing a "technological due process" model to vindicate procedural values in an era of automation); Kate Crawford & Jason Schultz, Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms, 55 B.C. L. Rev. 93 (2014) (arguing that procedural due process provides a framework for the regulation of big data); David Gray & Danielle Citron, The Right to Quantitative Privacy, 98 Minn. L. Rev. 62 (2013) (raising concerns over the use of algorithmic systems to establish probable cause for law enforcement searches or arrests).

[6] See Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, Machine Bias, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/BH93-NG7C].

to how many hours of care disabled individuals will receive,[7] to which employees should be hired, fired, or promoted.[8] Yet as decisionmaking shifts from human-only to a mixture of human and algorithm, questions of how to allocate constitutional liability have remained largely unanswered.

The majority of solutions to these concerns have focused on technological or regulatory oversight to address bias, fairness, and due process.[9] However, to date, few if any of these approaches have succeeded in providing adequate accountability frameworks, either because they have failed to address the larger social and structural aspects of the problems or because there is a lack of political will to implement them.[10] As such, it is time to consider new paradigms for accountability, especially for potential constitutional violations.

One underexplored approach is the possibility of holding AI vendors accountable for constitutional violations under the state action doctrine. Although state actors are typically governmental employees, a private party may be deemed a state actor if (1) the private party performs a function that is traditionally and exclusively performed by the state, (2) the state directs or compels the private party's conduct, or (3) the private party acts jointly with the government.[11]

This Essay explores this approach to AI accountability in three parts. Part I outlines the current state of play for government use of AI systems, especially those involved in key governmental decisionmaking processes. Part II reviews the relevant case law and literature on the state action

[7] See Colin Lecher, What Happens When an Algorithm Cuts Your Health Care, The Verge (Mar. 21, 2018), https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy [https://perma.cc/8SS7-F7K5]; see also infra section I.A.

[8] See Miranda Bogden & Aaron Rieke, Upturn, Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias 1–2 (2018), https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20–%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms,%20Equity%20and%20Bias.pdf [https://perma.cc/UQT6-4PSN]; Loren Larsen, HireVue Poised to Bring US Government Agencies' Recruiting Up to Speed, HireVue (May 16, 2019), https://www.hirevue.com/blog/hirevue-poised-to-bring-us-government-agencies-recruiting-up-to-speed [https://perma.cc/KWQ2-LW9H].

[9] See supra note 5.

[10] See, e.g., Bogden & Rieke, supra note 8, at 7 ("Structural kinds of bias also act as barriers to opportunity for jobseekers, especially when predictive tools are involved."); Dillon Reisman, Jason Schultz, Kate Crawford & Meredith Whittaker, AI Now Inst., Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability 3 (2018), https://ainowinstitute.org/aiareport2018.pdf [https://perma.cc/TE2M-HBUU] [hereinafter AI Now AIA Report] (proposing a comprehensive framework for assessing the "automated decision systems" of public agencies); Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz & Oscar Schwartz, AI Now Inst., AI Now 2018 Report 12 (2018), https://ainowinstitute.org/AI_Now_2018_Report.pdf [https://perma.cc/V7B8-4XPY] [hereinafter AI Now 2018 Report] ("[AI] tools . . . could easily be turned to more surveillant ends in the U.S., without public disclosure and oversight, depending on market incentives and political will.").

[11] Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1928 (2019); Sybalski v. Indep. Grp. Home Living Program, Inc., 546 F.3d 255, 257 (2d Cir. 2008).

doctrine, focusing on the public function, compulsion, and joint participation theories, and how these theories might apply to vendors of AI systems that government uses. Finally, Part III discusses the normative arguments in favor of applying the state action doctrine to close the AI accountability gap. Specifically, this Essay argues that—unlike traditional technology vendors that supply government actors with primarily functional tools, such as a computer operating system, word processing program, or web browser—AI vendors provide government with tools that assist or supply the core logic, justification, or action that is the source of the constitutional harm. Thus, much like other private parties whose conduct is fairly attributable to the state, vendors who build AI systems may also subject themselves to constitutional liability.

I. SEEING LIKE A STATE AI SYSTEM

To date, there is no comprehensive map or even agreed-upon methodology for tracking government use of AI in the United States, although some efforts are currently underway at the federal, state, and city levels.[12] Until the existence, design, and functions of these systems can be successfully documented, assessing their impact on constitutional accountability will be challenging.[13] In particular, there have been two main challenges to public scrutiny of AI: (1) lack of clear public accountability and oversight processes; and (2) objections from vendors that any real insights into their technology would reveal trade secrets or other confidential information.[14]

[12] For some preliminary attempts, see, e.g., Frank Pasquale, The Black Box Society: The Secret Algorithms that Control Money and Information 18, 140–58 (2015) (explaining that meaningful reform "mean[s] focusing less on trying to control the collection of data up front, and more on its *use*—how companies and governments are actually deploying it"); AI Now AIA Report, supra note 10, at 16 (proposing a self-assessment process that provides "an opportunity for agencies to develop expertise when commissioning and purchasing automated decision systems"); Robert Brauneis & Ellen P. Goodman, Algorithmic Transparency for the Smart City, 20 Yale J.L. & Tech. 103, 109 (2018) (testing the opacity of six predictive algorithms used by different local governments).

[13] See, e.g., AI Now AIA Report, supra note 10, at 6 (explaining that algorithmic impact assessments can "provide communities with information that can help determine whether those systems are appropriate"); Press Release, Admin. Conference of the U.S., ACUS Announces New Initiatives on the Use of Artificial Intelligence in the Federal Administrative Process (Nov. 28, 2018), https://www.acus.gov/newsroom/news/acus-announces-new-initiatives-use-artificial-intelligence-federal-administrative [https://perma.cc/4UUB-YECL] (announcing a collaboration with leading scholars to produce a report on AI's application to administrative adjudication and rulemaking); Automated Decision Systems: Examples of Government Use Cases, AI Now Inst. 1, https://ainowinstitute.org/nycadschart.pdf [https://perma.cc/YYW3-7YBT] (last visited July 29, 2019) (listing examples of government use of automated decision systems in New York City "to help New Yorkers understand the scope of issues and use cases").

[14] See, e.g., Brief for the AI Now Institute et al. as Amici Curiae Supporting the Respondent at 9–32, Food Mktg. Inst. v. Argus Leader Media, 139 S. Ct. 2356 (2019) (No. 18-481), 2019 WL 1453518 [hereinafter Brief for the AI Now Institute]; Hannah

Thus, while some information can be gleaned via public processes, investigative reporting, or open records, this information is often generalized and lacking in useful detail.[15] For example, at the federal level, the few glimpses into the state of AI have come through self-initiated processes, such as the Obama Administration's AI policy process.[16] Recently, the Administrative Conference of the United States announced its own process and forthcoming report on "existing and potential uses of artificial intelligence to improve administrative adjudication, rulemaking, and other regulatory activities throughout the federal government."[17] At the state and local levels, several commissions and task forces have begun to look into these questions.[18]

Bloch-Wehba, Access to Algorithms, 88 Fordham L. Rev. (forthcoming) (manuscript at 4–5), https://ssrn.com/abstract=3355776 (on file with the *Columbia Law Review*); Citron & Pasquale, supra note 5, at 5, 8; Natalie Ram, Innovating Criminal Justice, 112 Nw. U. L. Rev. 659, 663 (2018); Rebecca Wexler, Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System, 70 Stan. L. Rev. 1343, 1349–50 (2018).

[15] See, e.g., AI Now AIA Report, supra note 10, at 17 (explaining the difficulties involved with open record requests that arise because agencies do not necessarily know which data sets are pertinent); Brauneis & Goodman, supra note 12, at 152 ("Our research suggested that governments simply did not have many records concerning the creation and implementation of algorithms, either because those records were never generated or because they were generated by contractors and never provided to the governmental clients.").

[16] See Ed Felten & Terah Lyons, The Administration's Report on the Future of Artificial Intelligence, The White House: President Barack Obama (Oct. 12, 2016), https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence [https://perma.cc/73QA-N8FV]; see also Comm. on Tech., Exec. Office of the President, Preparing for the Future of Artificial Intelligence 1 (2016), https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf [https://perma.cc/8WJ5-MVK9]; Networking & Info. Tech. Research & Dev. Subcomm., Exec. Office of the President, The National Artificial Intelligence Research and Development Strategic Plan 3 (2016), https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf [https://perma.cc/2XSC-8774]. Note that one of the authors of this Essay, Jason Schultz, was a member of the Office of Science and Technology Policy AI Policy Team during this period and contributed to some of the Team's reports. The Trump Administration has also flagged some of these issues in the President's recent executive order on artificial intelligence. See Exec. Order No. 13,859, 84 Fed. Reg. 3967 (Feb. 11, 2019). However, it is unclear exactly what efforts will be made as the order does not provide any direction, resources, or other guidance on how to proceed.

[17] Admin. Conference of the U.S., supra note 13.

[18] See, e.g., Idaho Code § 19-1910 (2019) (establishing that all pretrial risk assessment tools should be transparent and open to public assessment); DJ Pangburn, How to Lift the Veil off Hidden Algorithms, Fast Company (Jan. 28, 2019), https://www.fastcompany.com/90292210/transparency-government-software-algorithms [https://perma.cc/2DAN-G4EV] (noting algorithmic accountability efforts in Washington State; Santa Clara County, California; Oakland, California; Berkeley, California; New York City; and Seattle); New York City Automated Decisions Task Force, NYC.gov, https://www1.nyc.gov/site/adstaskforce/index.page [https://perma.cc/64H5-PNDM] (last visited July 29, 2019) (describing the composition and purpose of New York City's task force to evaluate the implications of automation in the city's decisionmaking).

Behind the scenes, however, government use of privately designed algorithmic systems is increasing. For example, there have been reports—but little meaningful public disclosure—on the development of various systems within Immigration and Customs Enforcement that raise constitutional concerns, including at least one system provided by the well-known data analytics firm Palantir.[19] In particular, evidence that has recently come to light suggests that Palantir helped provide the "intelligence" to identify and separate undocumented immigrant children from their families.[20] Another example is the Trump Administration's recent budget request to Congress, which included a section from the Social Security Administration (SSA) explaining that it would study "whether to expand the use of social media networks in disability determinations, partly to help identify fraud."[21] While the current budget proposal discusses human monitoring,[22] a separate SSA document notes that it is developing an "Anti-Fraud Enterprise Solution," which will "integrate data from multiple sources and use industry-proven predictive analytics software to identify high-risk transactions for further review."[23] The

[19] See Nat'l Immigration Project of the Nat'l Lawyers Guild, Immigrant Def. Project & Mijente with Empower, LLC & Ford Found., Who's Behind ICE? The Tech and Data Companies Fueling Deportations 31–35, 38, 43–45 (2018), https://mijente.net/wp-content/uploads/2018/10/WHO'S-BEHIND-ICE_-The-Tech-and-Data-Companies-Fueling-Deportations_v3-.pdf [https://perma.cc/WN7H-2YVH] ("Palantir's new Integrated Case Management (ICM) system for ICE plays a key role in this information sharing with law enforcement."); Manish Singh, Palantir's Software Was Used for Deportation, Documents Show, TechCrunch (May 3, 2019), https://techcrunch.com/2019/05/03/palantirs-software-was-used-for-deportations-documents-show/ [https://perma.cc/Q8YE-SKRY]; Spencer Woodman, Palantir Provides the Engine for Donald Trump's Deportation Machine, The Intercept (Mar. 2, 2017), https://theintercept.com/2017/03/02/palantir-provides-the-engine-for-donald-trumps-deportation-machine/ [https://perma.cc/YYA9-HNAX].

[20] See April Glaser, Palantir Said It Had Nothing to Do with ICE Deportations. New Documents Seem to Tell a Different Story., Slate (May 2, 2019), https://slate.com/technology/2019/05/documents-reveal-palantir-software-is-used-for-ice-deportations.html [https://perma.cc/9Y9B-RUQ9].

[21] Robert Pear, On Disability and on Facebook? Uncle Sam Wants to Watch What You Post, N.Y. Times (Mar. 10, 2019), https://www.nytimes.com/2019/03/10/us/politics/social-security-disability-trump-facebook.html (on file with the *Columbia Law Review*).

[22] See Soc. Sec. Admin., Fiscal Year 2020 Budget Overview 26 (2019), https://www.ssa.gov/budget/FY20Files/2020BO_1.pdf [https://perma.cc/9R42-LYL8] ("In FY 2019, we are evaluating how social media could be used by disability adjudicators in assessing the consistency and supportability of evidence in a claimant's case file.").

[23] Soc. Sec. Admin., Annual Performance Report: Fiscal Years 2017–2019, at 26 (2018), https://www.ssa.gov/budget/FY19Files/2019APR.pdf [https://perma.cc/RC88-RPTM]. For a discussion of government use of AI in the child welfare context, see Virginia Eubanks, Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor 127–73 (2017) (discussing the use of the Allegheny Family Screening Tool to "forecast child abuse and neglect" for child welfare decisions); Dan Hurley, Can an Algorithm Tell When Kids Are in Danger?, N.Y. Times Mag. (Jan. 2, 2018), https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html (on file with the *Columbia Law Reivew*) ("Allegheny's Family Screening Tool is drawing interest from child-protection agencies around the country.").

Trump Administration's 2019 executive order on artificial intelligence also lays the path for all federal agencies to adopt AI systems moving forward.[24]

As we slowly learn more about these systems, it is becoming clear that they represent a range of public–private configurations. Some are developed entirely "in-house" by government,[25] others by contractors or as a licensed service,[26] or even as a "donation,"[27] which may impede oversight. But few publicly available documents note specifically how constitutional accountability is allocated in each system, especially within joint public–private endeavors.

[24] See Cade Metz, Trump Signs Executive Order Promoting Artificial Intelligence, N.Y. Times (Feb. 11, 2019), https://www.nytimes.com/2019/02/11/business/ai-artificial-intelligence-trump.html (on file with the *Columbia Law Review*) ("The administration . . . will call on government agencies to develop fellowships related to A.I.").

[25] Craig McCarthy, NYPD Uses New Tool to Find Crime Patterns: Officials, N.Y. Post (Mar. 10, 2019), https://nypost.com/2019/03/10/nypd-uses-new-tool-to-find-crime-patterns-officials/ [https://perma.cc/9MSH-W8FP] ("The software, designed in-house over two years and dubbed Patternizr, automates traditional police legwork through machine learning to find patterns in crimes.").

[26] See, e.g., William Alden, There's a Fight Brewing Between the NYPD and Silicon Valley's Palantir, BuzzFeed News (June 28, 2017), https://www.buzzfeednews.com/article/williamalden/theres-a-fight-brewing-between-the-nypd-and-silicon-valley [https://perma.cc/R95V-LXTH] (describing the NYPD's software contract with Palantir); Rich Duprey, Why I Think Axon Enterprise's Q4 Earnings Miss Is a Gift to Investors, Motley Fool (Mar. 4, 2019), https://www.fool.com/investing/2019/03/04/why-i-think-axon-enterprises-q4-earnings-miss-is-a.aspx [https://perma.cc/76KU-FY9A] (last updated Apr. 10, 2019) (explaining how law enforcement agencies use Axon's software to manage and analyze data from body cameras); Clare Garvie, Alvaro Bedoya & Jonathan Frankle, The Perpetual Line-Up: Unregulated Police Face Recognition in America, Georgetown Law Ctr. on Privacy & Tech. (Oct. 18, 2016), https://www.perpetuallineup.org [https://perma.cc/G76J-DLZL] (reporting that some police departments have purchased facial recognition software from third parties); George Joseph & Kenneth Lipp, IBM Used NYPD Surveillance Footage to Develop Technology that Lets Police Search by Skin Color, The Intercept (Sept. 6, 2018), https://theintercept.com/2018/09/06/nypd-surveillance-camera-skin-tone-search/ [https://perma.cc/W9JE-Z43W] (describing the NYPD's use of IBM video surveillance software "to identify suspicious objects and persons"); Lecher, supra note 7 (explaining that Arkansas used a third-party vendor to develop an algorithm to assess eligibility for disability benefits); Daniela Silva & Cyrus Farivar, ACLU Calls for U.S. Law Enforcement to Stop Sharing License Plate Data with ICE, NBC News (Mar. 13, 2019), https://www.nbcnews.com/tech/security/aclu-calls-u-s-law-enforcement-stop-sharing-license-plate-n983021 [https://perma.cc/HH5N-3NGM] (noting that ICE contracts with a vendor to access data from automated license plate readers).

[27] See, e.g., Michelle Chen, Beware of Big Philanthropy's New Enthusiasm for Criminal Justice Reform, Nation (Mar. 16, 2018), https://www.thenation.com/article/beware-of-big-philanthropys-new-enthusiasm-for-criminal-justice-reform/ [https://perma.cc/UCS4-TRAT] (discussing the Koch brothers' philanthropic network's "Safe Streets and Second Chances" initiative for reducing recidivism using "technology-driven reentry programs"); Ali Winston, New Orleans Ends Its Palantir Predictive Policing Program, The Verge (Mar. 15, 2018), https://www.theverge.com/2018/3/15/17126174/new-orleans-palantir-predictive-policing-program-end [https://perma.cc/AHM6-4V3X] (reporting that the New Orleans Police Department had a pro bono contract to use Palantir's software "to identify potential aggressors and victims of violence").

A. *Litigating AI Accountability: Four Case Studies*

In mapping AI accountability, one often overlooked resource is the courtroom, "where evidence, expert testimony, and judicial scrutiny reveal new insights into the current state of AI systems."[28] Recently, we spearheaded an effort via the AI Now Institute, in partnership with the New York University School of Law's Center on Race, Inequality, and the Law and the Electronic Frontier Foundation, "to conduct an examination of current United States courtroom litigation where the use of algorithms by government was central to the rights and liberties at issue in the case."[29] Our first report focused primarily on three case studies in which AI decisionmaking systems were already prevalent in government: (1) Medicaid and disability benefits,[30] (2) public teacher employment evaluations,[31] and (3) criminal risk assessment.[32] At our most recent workshop, we learned about new litigation involving the use of algorithmic decision systems in unemployment benefits.[33]

1. *Medicaid and Disability Benefits.* — Our first session began with the story of Tammy Dobbs, who has cerebral palsy. In 2008, Tammy moved from Missouri to Arkansas, where she was able to sign up for a state Medicaid waiver program to pay for a caretaker.[34] Tammy uses a wheelchair and does not have full use of her hands, so she needs help with many basic daily tasks, such as going to the bathroom and bathing.[35] The initial nurse that assessed Tammy under the program decided that she should have fifty-six hours of home care per week, the maximum allowable.[36] This all changed in 2016 when Tammy's annual assessment came with a new decisionmaker—an algorithm on a laptop computer.[37] The human assessor asked similar questions to those asked in previous years, but after entering the answers into the algorithmic system, Tammy's allowable care hours were reduced to thirty-two hours per week from fifty-six.[38] Worse yet, the computer system provided no explanation or opportunity to discuss the change, let alone guidance on how Tammy would be able to adjust to the reduction in home care hours.[39] The

[28] Litigating Algorithms Announcement, supra note 1.

[29] Id.; see also Litigating Algorithms, supra note 4, at 3.

[30] Litigating Algorithms, supra note 4, at 7–9.

[31] Id. at 10.

[32] Id. at 13–14.

[33] See Rashida Richardson, Jason M. Schultz & Vincent M. Southerland, AI Now Inst., Litigating Algorithms Report 2019 US Report (2019), https://ainowinstitute.org/litigatingalgorithms-2019-us.pdf [https://perma.cc/8GAD-3WB3] [hereinafter Litigating Algorithms 2019].

[34] See Lecher, supra note 7.

[35] Id.

[36] Id.

[37] Id.

[38] Id.

[39] See id.

human who accompanied the computerized decisionmaker couldn't help either, nor would the state when Tammy complained.[40]

Unfortunately, the story of Tammy Dobbs was not an isolated one. Hundreds of disabled Arkansans saw their care hours suddenly and drastically cut. They began to complain to the state, and later to attorneys, such as Legal Aid's Kevin De Liban, about the systematic and unexplained cuts in their disability benefits.[41] These cuts were all apparently the result of a new algorithmic system—an early form of AI—that the state had adopted as a cost-saving measure in an era when budgets were tight and healthcare costs continued to rise.[42] And Arkansas wasn't alone. A similar situation had also developed in Idaho under their Medicaid program.[43] Ultimately, both De Liban and the ACLU of Idaho sued in their respective states, claiming that the "faulty algorithmic decision systems improperly diminished or terminated benefits and services to individuals with intellectual, developmental, and physical disabilities."[44] The plaintiffs prevailed in both of these lawsuits, using a combination of constitutional and statutory claims to enjoin the use of these programs.[45] The failure of the state to provide adequate notice of the change—or an explanation of how the new algorithmic system would work—was central

[40] See id.

[41] See id.

[42] See id. (explaining that the "state ha[d] been prompted to look for new ways to contain costs and distribute what resources they ha[d]" available for healthcare).

[43] See id. ("[T]he state [of Idaho] made an attempt, like Arkansas, to institute an algorithm for allocating home care and community integration funds."); see also Leo Morales, Federal Court Rules Against Idaho Department of Health and Welfare in Medicaid Class Action, ACLU Idaho (Mar. 30, 2016), https://acluidaho.org/en/news/federal-court-rules-against-idaho-department-health-and-welfare-medicaid-class-action [https://perma.cc/CNA8-MVWS] (discussing a case involving the Idaho Department of Health and Welfare's automated decision system).

[44] Litigating Algorithms, supra note 4, at 7; see also K.W. ex rel. D.W. v. Armstrong, 180 F. Supp. 3d 703, 718 (D. Idaho 2016) (finding that the algorithmic system used to calculate Medicaid benefits violated due process); Ark. Dep't of Human Servs. v. Ledgerwood, 530 S.W.3d 336, 342–43 (Ark. 2017) (upholding an order enjoining the state from using its algorithm-based assessment tool to calculate attendant care hours, and finding irreparable harm to profoundly disabled beneficiaries). Disability rights advocates brought similar suits in West Virginia and Oregon. See Michael T. v. Bowling, No. 2:15-cv-09655, 2016 WL 4870284, at *1–4 (S.D. W. Va. Sept. 13, 2016); Order on Motion for Preliminary Injunction at 2, C.S. v. Saiki, No. 6:17-cv-00564-MC (D. Or. filed Apr. 19, 2017); Disability Rights Oregon Files Lawsuit About State Cuts in Home Care for Persons with Disabilities, City of Portland (Apr. 19, 2017), https://www.portlandoregon.gov/civic/article/635472 [https://perma.cc/93GS-BHCE].

[45] *K.W.*, 180 F. Supp. 3d at 719–20 (outlining the due process violation arising from the lack of notice given to beneficiaries regarding their benefit calculations); Ark. Dep't of Human Servs. v. Ledgerwood, 571 S.W.3d 911, 913 (Ark. 2019) (explaining that, on remand, the Arkansas circuit court issued a permanent injunction on the use of the challenged program); Class Action Settlement Agreement at 2–3, *K.W.*, 180 F. Supp. 3d 703 (No. 1:12-cv-00022-BLW) (outlining the plaintiffs' due process, equal protection, and statutory claims).

to each decision, rendering the deployment of the systems illegal.[46] Notably, the question of how to address the individualized deprivations for each plaintiff remained undetermined upon remand to the respective trial courts.[47]

These cases, however, were not situations in which negligent or vindictive government officials actively sought to deprive beneficiaries of their entitlements. Rather, these cases emerged when agencies attempted to implement complex yet archaic algorithmic formulas in computer systems to govern benefit assessment and disbursal. These AI systems were implemented without meaningful training, support, or oversight, and without any specific protections for recipients.[48] This was due in part to the fact that they were adopted to produce cost savings and standardization under a monolithic technology-procurement model, which rarely takes constitutional liability concerns into account.[49] Instead, "these systems typically target populations that are considered the 'most expensive,' which often include the most politically, socially, and economically marginalized communities, who, because of their status, are more likely to need greater levels of support."[50] Thus, an algorithmic system itself, optimized to cut costs without consideration of legal or policy concerns, created the core constitutional problems that ultimately decided the lawsuits.

These problems were also exacerbated as the result of a pattern that has emerged in which AI systems are adopted from state to state through a pattern of software contractor migration, by which AI vendors—like traveling sales representatives—usher the system from one state to another, training it on one state's historical data and then applying it to the new population.[51] Through this migration, patterns of bias or

[46] *Ledgerwood*, 530 S.W.3d at 341; see also *K.W.*, 180 F. Supp. 3d at 720 ("Crucial here is that the notice provide the reasons for the budget reduction so that the participant can challenge the reduction, and this requires the IAP to explain what she relied upon.").

[47] See, e.g., *K.W.*, 180 F. Supp. 3d at 722 (explaining that "[t]here are simply too many questions to rule as a matter of law" on either party's motion for summary judgment regarding the plaintiffs' individual claims).

[48] See Litigating Algorithms, supra note 4, at 7; see also Class Action Settlement Agreement, supra note 45, at 6–9.

[49] See Litigating Algorithms, supra note 4, at 7.

[50] Id.

[51] See id. ("Many states simply pick an assessment tool used by another state, trained on that other state's historical data, and then apply it to the new population . . . ."). Furthermore, "there are frequent flaws and errors in how these assessment systems are implemented and how they calculate the need for care." Id. Government agencies adopting these systems commonly enter into contracts with third-party vendors that handle everything. See supra note 26. The agency, particularly frontline staff that are most familiar with the Medicaid program and its challenges, has little to no involvement in how the tool analyzes data and produces calculations. See Litigating Algorithms, supra note 4, at 7–8. Because these tools are often based on private systems licensed to government agencies, the design specifications and particularities of the technical system are considered trade secrets of the vendor and are not publicly available. Id. at 8.

discrimination can proliferate technologically outside of the actions or intentions of any individual state employee.[52]

In terms of litigation outcomes, a key finding was that these cases involved claims against the government agencies alone and not the third-party AI vendors.[53] In bringing their claims against traditional government actors, plaintiffs were able to succeed, in part, on constitutional due process and administrative law theories that challenged the lack of notice, explanation, and ability to comment on or contest the changes to public benefit systems.[54] This was especially relevant for the plaintiffs, who were individuals with intellectual or developmental disabilities. Notably, procedural due process claims were able to overcome some of the arguments that disclosure of the technical workings of the systems would violate trade secrecy laws, especially when central to the question of how various public benefits determinations were made.[55]

However, these victories offered the plaintiffs and their advocates neither accountability for the core violations they experienced nor any real sense of protection against future harms from similar AI systems. In these cases, the court was willing to rule against the government's use of an AI system when it was deployed without constitutionally proper notice or when it produced discriminatory or otherwise inaccurate determinations.[56] But the claims and the court's jurisdiction were limited solely to the government agency, which had little to no actual involvement in the design, training, implementation, or testing of the system. In a sense, the state was merely a shell to house unconstitutional activity, not the primary

[52] See Litigating Algorithms, supra note 4, at 7.

[53] See K.W. ex rel. D.W. v. Armstrong, 180 F. Supp. 3d 703, 706–07, 718 (D. Idaho 2016); Ark. Dep't of Human Servs. v. Ledgerwood, 530 S.W.3d 336, 339–40 (Ark. 2017).

[54] In one case, "a court found that the state's automated Medicaid budgeting system was so unreliable that it 'arbitrarily deprive[d] participants of their property rights and hence violate[d] due process.'" Brief for the AI Now Institute, supra note 14, at 19 (quoting *K.W.*, 180 F. Supp. 3d at 718). In the same case, the court found that the state's refusal to provide a manual for a disability scoring tool furnished by a private company frustrated patients' ability to appeal. *K.W.*, 180 F. Supp. 3d at 717.

[55] Another finding was the extent to which discovery of errors in the AI's software design or implementation was connected to direct constitutional liability. Such connections were predicated on having access to technical information about the system and access to experts who have the ability to review and interpret the system, both of which can be difficult to obtain. For example, in the Arkansas case, the AI system allocating home healthcare to Medicaid patients failed to accurately understand the care needs of patients with conditions like cerebral palsy or multiple sclerosis. See *Ledgerwood*, 530 S.W.3d at 339–40, 343. Yet this was only discovered during the course of litigation, and only after the system's code and its associated technical documentation had been carefully examined. See Letter from Kevin De Liban, Legal Aid of Ark., to Becky Murphy, Office of Policy Coordination & Promulgation 5–6 (July 31, 2018) (on file with the *Columbia Law Review*); Marci Manley, Working 4 You: A Formula for Care, Finding a Solution, KARK (Nov. 17, 2017), https://www.kark.com/news/working-4-you-a-formula-for-care-finding-a-solution-2/ [https://perma.cc/B7K5-URBK].

[56] See supra notes 44–46 and accompanying text.

actor responsible for perpetrating it. At the end of the day, the plaintiffs still had very little understanding of exactly how and why the AI system had reduced their benefits, and even less of an opportunity to hold accountable the private technology vendors who were primarily responsible for the harm. Constitutional accountability mechanisms in the courts inherently involve core judicial concepts such as access to the evidence of the harm[57] and invocation of the court's appropriate remedial and prophylactic powers.[58] In the Arkansas and Idaho litigation, as well as their sister cases throughout the country, constitutional accountability for the creators of the AI systems responsible for the harms has been entirely absent.

2. *Public Teacher Employment Evaluations.* — The second case study explored similar themes in one of the few successful cases challenging the use of proprietary algorithms to evaluate the performance of public employees. In this case, a school district in Texas implemented a "data-driven" teacher-evaluation model through privately developed third-party software that purported to compare the results of a teacher's students to classroom statistics across the state and within the teacher's prior performance record.[59] The teachers sued the district through their union, arguing that the software was fundamentally inscrutable and that there was no way for teachers to know whether the software was accurately assessing their job performance.[60] The court agreed, holding that the "teachers have no meaningful way to ensure correct calculation of their [evaluation] scores, and as a result are unfairly subject to mistaken deprivation of constitutionally protected property interests in their jobs."[61] The court based its holding on procedural due process, finding that the teachers could proceed to trial on this constitutional issue.[62] The school district soon settled the case and stopped using the software.[63]

[57] See Crawford & Schultz, supra note 5, at 116.

[58] See Marbury v. Madison, 5 U.S. (1 Cranch) 137, 147 (1803) (explaining the "settled and invariable principle" that "every right, when withheld, must have a remedy").

[59] See Hous. Fed'n of Teachers, Local 2415 v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168, 1171–72 (S.D. Tex. 2017) ("In 2010, HISD began its transition to a 'data driven' teacher appraisal system . . . . The focus of this litigation is on the third criterion, student performance, particularly HISD's new method of rating teacher effectiveness based on proprietary algorithms belonging to a private company.").

[60] See id. at 1171.

[61] Id. at 1180.

[62] See id. at 1183 (denying summary judgment as to the plaintiffs' procedural due process claim).

[63] See Press Release, Am. Fed'n of Teachers, Federal Suit Settlement: End of Value-Added Measures for Teacher Termination in Houston (Oct. 10, 2017), https://www.aft.org/press-release/federal-suit-settlement-end-value-added-measures-teacher-termination-houston [https://perma.cc/3256-7N99]; see also Federal Lawsuit Settled Between Houston's Teacher Union and HISD, Hous. Pub. Media (Oct. 10, 2017), https://www.houstonpublicmedia.org/articles/news/2017/10/10/241724/federal-lawsuit-settled-between-houstons-teacher-union-and-hisd/ [https://perma.cc/DF65-WNQF].

Again, this case demonstrates one of the challenges of litigating AI claims when the entire algorithmic process is under the control of private third parties. Here, the challenged action was even more remote from the state than in the disability benefits cases discussed above. In this case, the state did not even house the AI system; instead, the system was built, trained, housed, and maintained entirely by a third-party software company, SAS Institute, Inc.[64] SAS fought to keep its source code, training data, and design as secret as possible, initially refusing to let the plaintiffs' experts see any of it and ultimately agreeing only to allow one expert to review the system under extreme constraints: only in the vendor's company office, only on a vendor laptop, and only with a pad of paper and a pen for note-taking.[65] While this lack of access ultimately supported the procedural due process ruling in favor of the teachers against the state,[66] it failed to provide any accountability mechanism against SAS that might have allowed the union to challenge the broader substantive concerns in the case, such as the union's equal protection claim or the claim that the system's determinations were arbitrary.

3. *Criminal Risk Assessment.* — The third case study focused on a juvenile sentencing hearing in Washington, D.C., in which the presiding judge declined to admit evidence from a long-standing "Violence Risk Assessment" system that had not been properly validated.[67] While the risk assessment in this case had not been implemented as part of an AI system, judges and other state actors at all levels of the criminal justice system rely on algorithmic tools to make decisions about detention and release.[68] In the case discussed at the workshop, participants noted:

[64] See *Hous. Fed'n of Teachers*, 251 F. Supp. 3d at 1177–79; Plaintiffs' Original Complaint at 13, *Hous. Fed'n of Teachers*, 251 F. Supp. 3d 1168 (No. 4:14-cv-01189), 2014 WL 1724308.

[65] See Plaintiffs' Response to Defendant's Motion for Summary Judgment at 41–42, *Hous. Fed'n of Teachers*, 251 F. Supp. 3d 1168 (No. 4:14-cv-01189), 2016 WL 9504197; see also Dr. Jesse Rothstein's Sworn Declaration Under Penalty of Perjury Pursuant to 28 U.S.C. § 1746 at 59–60, *Hous. Fed'n of Teachers*, 251 F. Supp. 3d 1168 (No. 4:14-cv-01189).

[66] See *Hous. Fed'n of Teachers*, 251 F. Supp. 3d at 1175–80, 1183.

[67] See Litigating Algorithms, supra note 4, at 13–14.

[68] See Erin Harbinson, Understanding 'Risk Assessment' Tools, Bench & B. Minn., Aug. 2018, at 14, 14–16. These tools purport to predict the risk that an individual will require rehabilitative resources while on parole, commit another offense after conviction, pose a threat to public safety, or fail to appear in court. See, e.g., id. at 16; see also Christopher Slobogin, Principles of Risk Assessment: Sentencing and Policing, 15 Ohio St. J. Crim. L. 583, 593–94 (2018). They rely on actuarial techniques to make predictions based on analysis of historical data. See Megan T. Stevenson & Christopher Slobogin, Algorithmic Risk Assessments and the Double-Edged Sword of Youth, 96 Wash. U. L. Rev. 681, 688 (2018); see also Ram, supra note 14, at 685. The appeal of risk assessment algorithms lies in their promise to objectively classify the likelihood of recidivism or failure to appear. See Aziz Z. Huq, Racial Equity in Algorithmic Criminal Justice, 68 Duke L.J. 1043, 1047 (2019). Without sufficient transparency, there is no way for the public to know whether any faults exist in a piece of software the government is using. See State v. Loomis, 881 N.W.2d 749, 763 (Wis. 2016), cert. denied, 137 S. Ct. 2290 (2017) (noting that transparency,

> [T]hese violence risk assessment systems have a powerful influence over criminal sentencing outcomes, especially for children. . . . [A] "high risk" finding on one of these algorithmic assessments can result in [a juvenile offender] being sent to a psychiatric hospital or a secured detention facility, separating them from their family and drastically changing the course of their life. Moreover, young people often plead guilty to violent offenses on the condition that they will be eligible for probation rather than incarceration, if they comply with certain court requirements including algorithmic risk assessments. When the risk assessment produces a high risk score, that score changes the sentencing outcome and can remove probation from the menu of sentencing options the judge is willing to consider.
>
> In examining these systems, many advocates have raised significant concerns about embedded racial bias. For example, most assessment systems include several risk factors that function as proxies for race. One risk factor that is often used is "parental criminality" which, given the long and well-documented history of racial bias in law enforcement, including the overpolicing of communities of color, can easily skew "high risk" ratings on the basis of a proxy for race. "Community disorganization" is another influential risk factor if an individual lives in a neighborhood considered to be "violent" or near gang activity, which given the long and well-documented history of private and public housing discrimination, could skew "high risk" ratings on the basis of a proxy for race.[69]

Even though defense attorneys were able to convince the judge to find the risk assessment inadmissible in that case, the ruling has not barred that particular assessment system or others from being used in subsequent cases in that juvenile court or in other courts and law enforcement contexts across the country.[70] As numerous AI vendors continue to license these tools, they will continue to evade broader accountability in courtrooms if their systems must be challenged on a case-by-case basis and the remedy is limited to the exclusion of the tool from specific cases.

4. *Unemployment Benefits.* — A final case study comes from our most recent workshop, which was held in June 2019.[71] In 2013, Michigan governor Rick Snyder launched the Michigan Integrated Data Automated System (MiDAS), a $47 million attempt to utilize the state's vast internal

accuracy, and due process concerns require that "use of a COMPAS risk assessment must be subject to certain cautions").

[69] Litigating Algorithms, supra note 4, at 13.

[70] See Beth Schwartzapfel, Can Racist Algorithms Be Fixed?, The Marshall Project (July 1, 2019), https://www.themarshallproject.org/2019/07/01/can-racist-algorithms-be-fixed [https://perma.cc/KN46-J77Q] (noting that some criminal justice advocacy groups encourage judges to use risk assessment algorithms "in context—as part of a larger decision-making framework that's sensitive to issues of racial justice").

[71] Litigating Algorithms 2019, supra note 33.

databases to detect and "robo-determin[e]" findings of fraud among recipients of unemployment benefits.[72] Specifically, after cross-checking data with employers, other state agencies, and the federal government, MiDAS "searched for discrepancies in the records of unemployment compensation recipients" and, if it found any, alerted the state Unemployment Insurance Agency (UIA) so that the claimant's file would be flagged as a potential case of misrepresentation.[73] When a file was flagged, MiDAS would initiate an automated process that attempted to transmit a multiple-choice questionnaire to the claimant, requiring a response within ten days.[74] However, because of the system's configuration, many questionnaires never arrived.[75] Others went to dormant accounts or to accounts of individuals whose benefits had already expired.[76]

> The questionnaire attempted to ask the recipient the following:
>
> Did you intentionally provide false information to obtain benefits you were not entitle[d] to receive?
>
> Yes No
>
> Why did you believe you were entitled to benefits?
>
> 1. I needed the money
> 2. I had not received payment when I reported for benefits
> 3. I reported the net dollar amount instead of the gross dollar amount paid
> 4. I did not understand how to report my earnings or separation reason
> 5. I thought my employer reported my earnings for me
> 6. Someone else certified (reported) for me
> 7. Someone else filed my claim for me
> 8. Other[77]

The system did not provide any other means of notice or response, and failure to respond to the questionnaire or *any* affirmative answer to even one question would result in a default determination that "the claimant knowingly and intentionally misrepresented or concealed information

[72] See Ryan Felton, Michigan Unemployment Agency Made 20,000 False Fraud Accusations—Report, Guardian (Dec. 18, 2016), https://www.theguardian.com/us-news/2016/dec/18/michigan-unemployment-agency-fraud-accusations [https://perma.cc/KD39-KUYE]; see also Mich. Office of the Auditor Gen., Performance Audit Report: Michigan Integrated Data Automated System (MiDAS) 29 (2016), https://audgen.michigan.gov/wp-content/uploads/2016/07/r641059315.pdf [https://perma.cc/QXL4-JWT8].

[73] See Cahoo v. SAS Inst. Inc. (*Cahoo II*), 377 F. Supp. 3d 769, 771–72 (E.D. Mich. 2019) ("MiDAS's electronic 'cross-checking' mechanism alerted the UIA when income was reported for claimants or when some activity affected a claimant's eligibility for benefits.").

[74] Id. at 772.

[75] Id.

[76] Id.

[77] Cahoo v. SAS Analytics Inc., 912 F.3d 887, 893 (6th Cir. 2019).

to unlawfully receive benefits."[78] Once the default determination was made, the UIA combined the MiDAS determination with its Enterprise Fraud Detection Software (provided by third-party vendor SAS Institute, Inc.) and sent the claimant a letter demanding repayment and assessing penalties plus interest, without any opportunity to appeal or otherwise contest the finding.[79] The penalties for nonpayment included "interception of the claimant's state and federal income tax refunds, garnishment of wages, and legal collection activity."[80]

Unfortunately for Snyder, the State of Michigan, and many of the recipients of its unemployment benefits, the system adjudicated over 22,000 fraud determinations with an error rate of 93%.[81] According to Steve Gray, the former director of Michigan Law's Unemployment Insurance Clinic and the current head of the UIA, those wrongly accused of fraud were subjected to "highest-in-the-nation quadruple penalties" and often weren't given sufficient notice of the adjudications to allow for proper appeals before the thirty-day deadline.[82] This resulted in an estimated tens of thousands of dollars per person in penalties, interest, and lost wages.[83] A subsequent class action lawsuit was brought against the State of Michigan over MiDAS on behalf of the class of recipients who had been wrongly accused.[84] In that case, the court found that the flaws in the MiDAS system had damaged plaintiffs and "eventually approved a settlement agreement in which the State agreed, among other things, to suspend all [MiDAS] collection activity."[85]

[78] Id. (quoting Plaintiff's First Amended Class Action Complaint and Jury Demand at 17–18, Cahoo v. SAS Inst. Inc., 322 F. Supp. 3d 772 (E.D. Mich. 2018) (No. 17-10657), 2017 WL 3405195). It is worth noting that this "matching" approach is quite similar to those that were used in Florida and Georgia to attempt to detect "voter fraud" with an "exact match" data-driven system prior to the 2018 elections. See generally Complaint, Ga. Coal. for the Peoples' Agenda, Inc. v. Kemp, 347 F. Supp. 3d 1251 (N.D. Ga. 2018) (No. 1:18-cv-04727-ELR) (alleging that Georgia's "Enet" system rejected registrations for voters if there was any mismatch between records on file with the Georgia Department of Drivers Services or Social Security Administration, even if the mismatch resulted from a government employee's typographical error during data entry).

[79] *Cahoo II*, 377 F. Supp. 3d at 772.

[80] Cahoo v. SAS Inst. Inc. (*Cahoo I*), 322 F. Supp. 3d at 786.

[81] Urging Your Support of the Bipartisan Legislative Package Reforming Michigan's Unemployment Insurance System: Hearing Before the S. Oversight Comm., 2017 Leg., 99th Sess. 2 (Mich. 2017) (testimony of Steve Gray, Clinical Assistant Professor and Director, Unemployment Insurance Clinic, University of Michigan Law School), http://www.senate.mi.gov/committeeMinTestimony/2017-2018/Oversight/Testimony/2017-SCT-OVER-11-30-1-05.PDF [https://perma.cc/V7CV-BV9Z].

[82] Id. at 2.

[83] Id.

[84] See Zynda v. Arwood, 175 F. Supp. 3d 791, 796–97 (E.D. Mich. 2016).

[85] *Cahoo I*, 322 F. Supp. 3d at 784; see also Stipulated Order of Dismissal at 5, UAW v. Arwood, No. 2:15-cv-11449 (E.D. Mich. Feb. 2, 2017), https://www.courtlistener.com/recap/gov.uscourts.mied.300638/gov.uscourts.mied.300638.51.0_1.pdf [https://perma.cc/5P4E-Q6FP].

However, as that case only addressed prospective relief against the state, plaintiffs brought a second class action lawsuit against both the individual state actors responsible for directly operating the system and the software companies that designed and implemented it for the state.[86] In that case, *Cahoo v. SAS Institute Inc.*, the complaint alleged that three separate technology vendors—FAST Enterprises, LLC; SAS Institute, Inc.; and CSG Government Solutions—"designed, created, implemented, or maintained the automated system employed by the UIA in adjudicating fraud determinations."[87] The plaintiffs alleged that these companies were state actors and thus were liable under 42 U.S.C. § 1983 for "the deprivation of a right secured by the Constitution or laws of the United States . . . caused by a person acting under the color of state law."[88]

In its decision on the defendants' motions to dismiss, the court specifically held that "these contracted companies and individuals, working alongside state officials, played some role in implementing a defective system that placed a significant financial burden on unemployment beneficiaries, and they acted under color of state law when doing so."[89] In other words, at least at the motion to dismiss stage of the civil case against them, the court found that the AI vendors were state actors. Exactly how and why is discussed in further detail below.[90]

II. The State Action Doctrine: A Framework for Private Actor Constitutional Accountability

While the above case studies show that pathways exist, to some degree, for holding governments accountable for how they use AI systems, they also highlight the stark fact that, until the *Cahoo* case, none of the third-party AI providers faced any liability for the constitutional harms their technology imposed. As shown below, this is largely because constitutional liability doctrines, including liability under 42 U.S.C. § 1983, have traditionally focused on the activities of public actors, such as government agencies or officials.[91] These doctrines operate under the assumption that government actors have both the greatest power and responsibility for upholding those rights and protections, and should therefore be held to the highest levels of accountability.[92] Meanwhile, private actors, such as corporations or citizens, need only be held accountable under traditional tort or regulatory approaches.[93] Or, as one

[86] See *Cahoo I*, 322 F. Supp. 3d at 787–88.

[87] Id. at 787.

[88] Id. at 791 (citing Dominguez v. Corr. Med. Servs., 555 F.3d 543, 549 (6th Cir. 2009)).

[89] Id. at 784.

[90] See infra Part II.

[91] Lillian BeVier & John Harrison, The State Action Principle and Its Critics, 96 Va. L. Rev. 1767, 1786 (2010) ("Constitutional rules are almost all addressed to the government.").

[92] See id. at 1794–96.

[93] See id. at 1794–97.

scholar puts it, "governmental power is, in general, more to be feared than nongovernmental power."[94]

However, when private actors wield the power of the state, or "act under color of state law," courts have sought to hold them as accountable as the state.[95] The state action doctrine mediates the border between private actors whose conduct is "fairly attributable to the state" and those whose conduct is seen as unrelated or external—a distinction that, despite its theoretical and formalistic dichotomy, has become increasingly difficult to maintain, if it even existed to begin with.[96] In particular, historical attempts to arbitrage constitutional protections through private sector "outsourcing" and the complex intertwining of public–private partnerships in the modern economy have challenged this separation as a sensible division in many arenas.[97] This is particularly problematic for AI systems, as the power and responsibility bestowed upon AI vendors to provide the functions of government is increasing dramatically. Thus, the applicability of the state action doctrine to AI vendors and their systems will be a central question for AI accountability going forward.[98]

To deal with this legacy and complexity, courts have been forced to evolve in their interpretation of the state action doctrine.[99] To assess

[94] Frederick Schauer, Acts, Omissions, and Constitutionalism, 105 Ethics 916, 916–17 (1995).

[95] See, e.g., Lugar v. Edmondson Oil Co., 457 U.S. 922, 942 (1982) ("In summary, petitioner was deprived of his property through state action; respondents were, therefore, acting under color of state law in participating in that deprivation.").

[96] See Gillian E. Metzger, Privatization as Delegation, 103 Colum. L. Rev. 1367, 1410–26 (2003) ("[I]t does not require a very robust or expansive understanding of government power in order to make the point that current state action doctrine is underinclusive.").

[97] See id. at 1400 ("Modern privatized government does not fit easily within the paradigms of U.S. constitutional law."); see also Martha Minow, Public and Private Partnerships: Accounting for the New Religion, 116 Harv. L. Rev. 1229, 1230 (2003) ("The new versions of privatization potentially jeopardize public purposes by pressing for market-style competition, by sidestepping norms that apply to public programs, and by eradicating the public identity of social efforts to meet human needs."); Neil Gordon, Contractors and the True Size of Government, POGO (Oct. 5, 2017), https://www.pogo.org/analysis/2017/10/contractors-and-true-size-of-government/ [https://perma.cc/7WSG-S9RF] (noting that "[m]ore than 40 percent of the [federal government] workforce—about 3.7 million people—are contract workers"); Lorraine Woellert & John Bresnahan, Sweeping Trump Proposal Seeks to Shrink Government, Merge Agencies, Politico (June 21, 2018), https://www.politico.com/story/2018/06/21/trump-shrink-federal-agencie-661976 [https://perma.cc/HYX3-XTSP] (explaining a proposal by the Trump Administration to privatize certain government entities such as Fannie Mae and Freddie Mac).

[98] While some vendors may voluntarily attempt to provide versions of transparency or accountability—either for internal ethical reasons or because of external market pressures—the lack of any legal accountability remains a concern, especially in situations in which government actors have little or no incentive to impose accountability on vendors through the contractual or procurement processes.

[99] See Edmonson v. Leesville Concrete Co., 500 U.S. 614, 620 (1991) ("Although the conduct of private parties lies beyond the Constitution's scope in most instances, governmental authority may dominate an activity to such an extent that its participants

constitutional liability for private parties under the state action doctrine, courts have generally applied three tests: (1) the public function test, which asks whether the private entity performed a function traditionally and exclusively performed by government;[100] (2) the compulsion test, which asks whether the state significantly encouraged or exercised coercive power over the private entity's actions;[101] and (3) the joint participation test, which asks whether the role of private actors was "pervasively entwined" with public institutions and officials.[102]

Despite this seemingly well-articulated approach, Supreme Court cases on the subject of state action have "not been a model of consistency,"[103] and therefore courts generally have "no single test to identify state actions and state actors."[104] Courts often look to "a host of facts that can bear on the fairness of an attribution of a challenged action to the State."[105] Thus, the fundamental question under each test is whether the

must be deemed to act with the authority of the government and, as a result, be subject to constitutional constraints."); Blum v. Yaretsky, 457 U.S. 991, 1002–05 (1982) ("[O]ur precedents indicate that a State normally can be held responsible for a private decision only when it has exercised coercive power or has provided such significant encouragement, either overt or covert, that the choice must in law be deemed to be that of the State."); *Lugar*, 457 U.S. at 936–37 (noting that "[i]t is a fundamental fact of our political order" that federal law limits constitutional liability to actions "fairly attributable to the State"). Note that "state action" has generally been found to be synonymous with "under color of law" in the Fourteenth Amendment context. See United States v. Price, 383 U.S. 787, 794 & n.7 (1966).

[100] See Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1933 (2019) (holding that a private nonprofit corporation designated by New York City to run a public access television channel was not a state actor bound by the First Amendment); Rendell-Baker v. Kohn, 457 U.S. 830, 835, 842 (1982) (applying the public function test in the context of the First, Fifth, and Fourteenth Amendments); Flagg Bros., Inc. v. Brooks, 436 U.S. 149, 158, 163 (1978) (noting that the public function test has "carefully confined bounds" because, "[w]hile many functions have been traditionally performed by governments, very few have been 'exclusively reserved to the State'"); Jackson v. Metro. Edison Co., 419 U.S. 345, 352–54 (1974) (refusing to expand the public function test to include all actions of any business that is "affected with the public interest").

[101] See Am. Mfrs. Mut. Ins. Co. v. Sullivan, 526 U.S. 40, 52–58 (1999) (holding that private insurers were not subject to constitutional liability when the state neither coerced nor encouraged the insurers' actions); *Rendell-Baker*, 457 U.S. at 841–43 (finding that a state-funded private school for children with special needs was not a state actor because there was no coercion or influence by the state on the challenged employment decision).

[102] See Brentwood Acad. v. Tenn. Secondary Sch. Athletic Ass'n, 531 U.S. 288, 298–302 (2001) (finding that a private school association was a state actor due to the "pervasive entwinement" of its activities with public institutions and officials).

[103] *Edmonson*, 500 U.S. at 632 (O'Connor, J., dissenting).

[104] *Brentwood*, 531 U.S. at 294.

[105] Cooper v. U.S. Postal Serv., 577 F.3d 479, 491 (2d Cir. 2009) (internal quotation marks omitted) (quoting Horvath v. Westport Library Ass'n, 362 F.3d 147, 151 (2d Cir. 2004)); see also *Jackson*, 419 U.S. at 349–50 ("[T]he question whether particular conduct is 'private,' on the one hand, or 'state action,' on the other, frequently admits of no easy answer.").

private entity's challenged actions are "fairly attributable" to the state.[106] Below, we examine each test to determine its applicability to AI vendors.

A. *The Public Function Test*

The first test focuses on whether the private actor is engaged in a core governmental function that has been exclusively and traditionally performed by the state.[107] As the Supreme Court noted this term in *Manhattan Community Access Corp. v. Halleck*, very few "functions" remain exclusive to the state in the modern era of public–private partnerships and competition.[108] Rather, many functions are shared or more diversely administered—such as "administering insurance payments, operating nursing homes, providing special education, . . . supplying electricity,"[109] or—as the majority found in *Halleck*—"operating public access channels on a cable system."[110]

When a traditional and exclusive public function is "outsourced" to a private entity, however, it still may fall within the scope of the state action doctrine's purview.[111] For example, in *West v. Atkins*, the Supreme Court unanimously held that a private medical provider that contracted to run a clinic for a North Carolina prison engaged in state action by treating inmates.[112] There, the Court ruled that "the State was constitutionally obligated to provide medical treatment to injured inmates, and the delegation of that traditionally exclusive public function to a private physician gave rise to a finding of state action."[113] In holding the private medical provider liable as a state actor, the Court reasoned that "[c]ontracting out prison medical care does not relieve the State of its constitutional duty to provide adequate medical treatment to those in its custody, and it does not deprive the State's prisoners of the means to

[106] *Rendell-Baker*, 457 U.S. at 838 (internal quotation marks omitted) (quoting Lugar v. Edmondson Oil Co., 457 U.S. 922, 937 (1982)); see also *Am. Mfrs. Mut. Ins. Co.*, 526 U.S. at 50 (requiring the consideration of "whether the allegedly unconstitutional conduct is fairly attributable to the State"); Burton v. Wilmington Parking Auth., 365 U.S. 715, 725–26 (1961) ("Owing to the very 'largeness' of government, a multitude of relationships might appear to some to fall within the Amendment's embrace, but that, it must be remembered, can be determined only in the framework of the peculiar facts or circumstances present.").

[107] Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1928–31 (2019); see also Flagg Bros., Inc. v. Brooks, 436 U.S. 149, 157–58 (1978) ("While many functions have been traditionally performed by governments, very few have been 'exclusively reserved to the State.'" (quoting *Jackson*, 419 U.S. at 352 (1974))).

[108] *Halleck*, 139 S. Ct. at 1928–31 ("Under the Court's cases, those functions include, for example, running elections and operating a company town.").

[109] Id. at 1929.

[110] Id. at 1930.

[111] Id. at 1929 n.1.

[112] 487 U.S. 42, 55–58 (1988).

[113] Am. Mfrs. Mut. Ins. Co. v. Sullivan, 526 U.S. 40, 55 (1999) (interpreting the *West* holding).

vindicate their Eighth Amendment rights."[114] Even Justice Scalia in his concurrence in part agreed that "a physician who acts on behalf of the State to provide needed medical attention to a person involuntarily in state custody (in prison or elsewhere) and prevented from otherwise obtaining it," is constitutionally liable when the physician "causes physical harm to such a person by deliberate indifference."[115]

Two key considerations emerge from this analysis. First, the exclusivity of the function is not defined by competition—there are many medical providers in the world—but rather by the voluntary ability of the plaintiff to access (or obtain) the benefits of that functionality elsewhere. In *West*, because the plaintiff was a prisoner of the state, no such alternative access existed.[116] Second, the completely private status of the actor in question has little to do with constitutional liability. Instead, it is the role that actor plays in the administration of the state's function that governs.

As a result, the public function theory often goes beyond the formalism of identifying functions by type and instead looks more deeply at the stakes of delegating the specific function at issue to private actors. For example, in *Giron v. Corrections Corp. of America*, the district court applied the doctrine to a private management company that ran a state prison.[117] There, the court reasoned that "[i]f a state government must satisfy certain constitutional obligations when carrying out its functions, it cannot avoid those obligations and deprive individuals of their constitutionally protected rights by delegating governmental functions to the private sector. . . . The delegation of the function must carry with it a delegation of constitutional responsibilities."[118] Similarly, in *DeBauche v. Trani*, the Fourth Circuit found that a private party is a state actor "when the state has sought to evade a clear constitutional duty through delegation to a private actor . . . [or] delegated a traditionally and exclusively public function to a private actor."[119] In fact, the Supreme Court later went on to frame the public function approach as one almost imposing a legal duty of care on the government in the context of incarceration, explaining in *DeShaney v. Winnebago County Department of Social Services* that "when the State takes a person into its custody and holds him there against his will, the Constitution imposes upon it a corresponding duty to assume some responsibility for his safety and general well-being."[120]

[114] *West*, 487 U.S. at 56.

[115] Id. at 58 (Scalia, J., concurring in part and concurring in the judgment).

[116] See id. at 43–44 (majority opinion).

[117] 14 F. Supp. 2d 1245, 1247–48 (D.N.M. 1998).

[118] Id. at 1250 (citing Terry v. Adams, 345 U.S. 461 (1953)).

[119] 191 F.3d 499, 507 (4th Cir. 1999).

[120] 489 U.S. 189, 199–200 (1989); see also *West*, 487 U.S. at 55–56 (explaining that it was "the physician's function within the state system" and not his nongovernmental status that drove the determination that his actions could "fairly be attributed to the State"); Shields v. Ill. Dep't of Corr., 746 F.3d 782, 797 (7th Cir. 2014) ("A business . . . that

Otherwise, the state would "be free to contract out all services which it is constitutionally obligated to provide and leave its citizens with no means for vindication of those rights."[121] This approach to constitutional responsibility has also been extended to private transportation companies that serve prisons, as well as private residential treatment centers for children under the state's custodial care.[122]

Such questions of function and responsibility played out in the recent *Halleck* case before the Supreme Court. Justice Kavanaugh, writing for the majority, framed the role of the private telecommunications corporation as merely providing and operating the forum of public access television, which is a nonexclusive function.[123] Justice Sotomayor, writing for the dissenters, argued that the corporation did much more than that, having accepted the state's delegation of responsibility for administration and decisionmaking that was so central to the function of a public forum that constitutional liability was appropriate.[124]

Thus, when private AI vendors provide their software to governments to fulfill duties that are specifically tied to a state's overall public and constitutional obligations, the possibility of the vendor being held a state actor becomes a reality. For example, when governments use privately provided AI systems to support determinations of criminal propensity[125] or child welfare interventions,[126] the traditional and exclusive nature of those functions, along with their constitutional obligations, puts the AI provider in a similar role to the physician in *West.* The key question, then, is whether AI vendors—and the systems they create—are merely tools that government employees use to perform state functions,

contracts to provide medical care to prisoners undertakes 'freely, and for consideration, responsibility for a specific portion of the state's overall [constitutional] obligation to provide medical care for incarcerated persons.'" (alteration in original) (quoting Rodriguez v. Plymouth Ambulance Serv., 577 F.3d 816, 827 (7th Cir. 2009))).

[121] *West,* 487 U.S. at 56 n.14 (internal quotation marks omitted) (quoting West v. Atkins, 815 F.2d 993, 998 (4th Cir. 1987) (Winter, C.J., concurring and dissenting)).

[122] See Lemoine v. New Horizons Ranch & Ctr., 990 F. Supp. 498, 501–02 (N.D. Tex. 1998) (finding that where the state provides housing, food, medical, and educational services to children involuntarily committed to private juvenile residential treatment centers, the owners and employees of those centers may be considered state actors, even when the state has delegated full authority to administer care to the private actor); see also Nguyen v. Prisoner Transp. Servs., No. 3:18-cv-00871, 2019 WL 429678, at *4 (M.D. Tenn. Feb. 4, 2019) (collecting cases of prison transport companies held to be state actors because they could not have performed those services without state authorization).

[123] See Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1929–31 (2019) ("[M]erely hosting speech by others is not a traditional, exclusive public function and does not alone transform private entities into state actors . . . .").

[124] See id. at 1940 (Sotomayor, J., dissenting) ("When a government (1) makes a choice that triggers constitutional obligations, and then (2) contracts out those constitutional responsibilities to a private entity, that entity—in agreeing to take on the job—becomes a state actor for purposes of § 1983.").

[125] The COMPAS system provides one example. See Angwin et al., supra note 6.

[126] See Hurley, supra note 23.

or whether the vendor systems perform the functions themselves.[127] If one views the provision of AI as simply the latest technological "tool"—like the provision of a hammer—to the state, then the private vendors are outside the definition of state action. On the other hand, when the purpose of the AI is to support or take on a role in the decisionmaking functions of a government official, one could easily imagine a court finding that work fitting within the public function test, more similar to the *West* example of the private medical professional using his professional judgment in deciding which medical services to provide as opposed to simply providing a scalpel or X-ray machine to a prison hospital.

In fact, many AI vendors specifically optimize their systems to attempt to approximate what a human actor would decide in a similar situation.[128] Take, for example, the Allegheny Family Screening Tool (AFST), an AI system that attempts to forecast child abuse and neglect so that child welfare workers can intervene preemptively and prevent "increased occurrences of drug and alcohol abuse, suicide attempts, and depression" among children in abusive or neglectful situations.[129] The tool, created by a team from Auckland University of Technology, purports to rank the danger of a child's situation from "a green 1 (lowest risk) at the bottom to a red 20 (highest risk) on top."[130] According to one report, AFST was based on a statistical analysis of prior child welfare calls, including "100 criteria maintained in eight databases for jails, psychiatric services, public-welfare benefits, drug and alcohol treatment centers."[131]

When vendors supply AI systems to government agencies, the results of the decisions and actions taken are not only attributable to the state but also effectuated through it. In particular, when AI systems are designed specifically for use in governmental domains, such as criminal justice, benefits determinations, or public employment, the conclusion that

[127] Note that we discuss AI systems here because, while vendors provide them currently, they may at some point become so sophisticated that constitutional liability may need to apply to these systems themselves. Of course, therein lies a problem with remedies. See Mark Lemley & Bryan Casey, Remedies for Robots 3 (Stanford Law & Econ. Olin Working Paper No. 523, 2018), https://ssrn.com/abstract=3223621 (on file with the *Columbia Law Review*) ("[I]t turns out to be much harder for a judge to 'order' a robot, rather than a human, to engage in or refrain from certain conduct.").

[128] See Ke Li, Learning to Optimize with Reinforcement Learning, Berkeley Artificial Intelligence Research (Sept. 12, 2017), https://bair.berkeley.edu/blog/2017/09/12/learning-to-optimize-with-rl/ [https://perma.cc/P2LT-8UHX] (describing optimization techniques that attempt to replicate in "concrete algorithms" the process by which "humans not only reason, but also reason about their own process of reasoning").

[129] Virginia Eubanks, A Child Abuse Prediction Model Fails Poor Families, WIRED (Jan. 15, 2018), https://www.wired.com/story/excerpt-from-automating-inequality/ [https://perma.cc/6QLL-HE69].

[130] Hurley, supra note 23.

[131] Id.

their design is a core public function is not difficult to imagine.[132] For example, the MiDAS and Enterprise Fraud Detection Software (EFDS) tools in the *Cahoo* case[133] were specifically designed, developed, and implemented to automate the determination of which unemployment beneficiaries to investigate and penalize, a set of decisions previously—and, after the litigation, subsequently—initiated, conducted, and supervised by government employees.[134]

All of the case studies above demonstrate situations in which algorithmic and AI systems are performing traditional public functions. In the disability cases, it is the function of assessing and recommending public benefit eligibility.[135] In the public employment context, it is the function of assessing and recommending human resources actions.[136] In the criminal justice context, it is the function of evaluating the dangerousness of a defendant.[137] And in the unemployment benefits context, it is the function of investigating and enforcing antifraud regulations.[138] Thus, in many contexts, the case for considering AI vendors as performing public functions could be quite strong.

B. *The Compulsion Theory*

The second test asks whether the action taken by private entities was encouraged, controlled, or compelled by the state, rather than being done with the "mere approval or acquiescence of the State,"[139] or, as it was continually framed in the recent oral argument in *Manhattan Community Access Corp. v. Halleck*, the extent to which the private entity has discretion to make substantive choices that impact constitutional concerns.[140]

For government use of AI, the determination would be quite fact dependent, but to the extent that any allegations of constitutional liability were based on inputs or designs given to the vendor by the state, this could qualify. For example, in *Cahoo*, the complaint alleged that CSG, the

[132] Note that this would not impact the provision of general-purpose software or even general-purpose AI to government agencies. Cf. Jackson v. Metro. Edison Co., 419 U.S. 345, 358–59 (1974) (finding private utility companies, even heavily regulated ones, not to be state actors).

[133] See infra sections II.B–.C for a discussion of the alternative state action theories addressed in *Cahoo.*

[134] See *Cahoo I*, 322 F. Supp. 3d 772, 785, 788 (E.D. Mich. 2018), aff'd in part and rev'd in part, 912 F.3d 887 (6th Cir. 2019).

[135] See supra section I.A.1.

[136] See supra section I.A.2.

[137] See supra section I.A.3.

[138] See supra section I.A.4.

[139] Am. Mfrs. Mut. Ins. Co. v. Sullivan, 526 U.S. 40, 52 (1999).

[140] See Transcript of Oral Argument at 46, Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921 (2019) (No. 17-1702), 2019 WL 2493920 ("If [the private party] has discretion so it can exercise editorial control, then it would not be a public forum.").

software company that ran and administered the UIA Project Control Office, was charged to do so by the State of Michigan and "received significant encouragement from the State when it implemented, configured, administered and maintained the defective and unconstitutional fraud detection system."[141] Moreover, the contract between CSG and the State of Michigan allegedly delegated managerial authority over the entire project to CSG.[142] In this sense, the state could be seen to have compelled CSG to take actions subject to constitutional liability.

In the other case studies, there are similar elements of compulsion. For example, in the *Houston Federation of Teachers* case, the key algorithmic inputs—student test scores—were provided entirely by the state, and the federal government required the "value-added" model upon which the AI system was based as a condition to receive $4.35 billion in Race to the Top funds.[143] In the disability benefits cases, much of the logic of the systems and the classification of conditions stems from mandatory state and federal regulations.[144] On the other hand, in criminal risk assessment cases in which the judge ultimately retains the discretionary authority to impose detention and other sentencing conditions, the doctrine might be less applicable.[145] Thus, for AI, the question of who controls the decisions for the design and implementation of the systems, including who provides the data to train and test the system, is relevant.[146]

[141] *Cahoo I*, 322 F. Supp. 3d 772, 793 (E.D. Mich. 2018), aff'd in part and rev'd in part, 912 F.3d 887 (6th Cir. 2019).

[142] Id.

[143] Audrey Amrein-Beardsley & Mark A. Paige, "Houston, We Have a Lawsuit:" A Cautionary Tale for the Implementation of Value-Added Models (VAMs) for High-Stakes Employment Decisions 1 (unpublished manuscript) (on file with the *Columbia Law Review*).

[144] See, e.g., Michael T. v. Bowling, No. 2:15-cv-09655, 2016 WL 4870284, at *1 (S.D. W. Va. Sept. 13, 2016) (explaining that, "[o]nce a state elects to provide an optional [Medicaid] service" like home-based disability care, "it must adhere to the pertinent federal statutes and regulations").

[145] See, e.g., State v. Loomis, 881 N.W.2d 749, 768 (Wis. 2016) (explaining that risk assessment tools are "merely one tool available to a court at the time of sentencing and a court is free to rely on portions of the assessment while rejecting other portions").

[146] See, e.g., Petition for Writ of Certiorari at 28, Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921 (2019) (No. 17-1702), 2018 WL 3129068 ("[B]ecause 'decisions regarding the programming on public access cable channels in the District of Columbia [are not alleged to] in any way [be] controlled by the District of Columbia government,' 'there is no state actor and thus no viable Section 1983 claim.'" (alterations in original) (quoting Glendora v. Sellers, No. 1:02-cv-00855, slip op. at 6 (D.D.C. Mar. 31, 2003))). This has also emerged as a basis for finding state action within the context of the Fourth Amendment and private party searches when the private party acts as "an agent or instrument of the [g]overnment." Gray & Citron, supra note 5, at 135–36 (internal quotation marks omitted) (quoting Skinner v. Ry. Labor Execs.' Ass'n, 489 U.S. 602, 614–15 (1989)).

C. *The Joint Participation Theory*

The third test for state action asks whether the government was significantly involved in the challenged action that is alleged to have caused the constitutional harm, so much so that the two entities can be considered joint participants.[147] For example, in *Brentwood Academy v. Tennessee Secondary School Athletic Ass'n*, the Supreme Court held that the private association was a state actor in part because a substantial majority of its members were public schools and public school officials "overwhelmingly perform all but the purely ministerial acts by which the Association exists and functions."[148] By contrast, in *Blum v. Yaretsky*, the Court considered "whether the decision of nursing homes to transfer or discharge patients constituted state action in light of the state's requirement that physicians certify the medical necessity of nursing home services on a 'long term care placement form' devised by the state."[149] The Court found no state action, holding that even though the state had created the form for evaluating patients, "the physicians, and not the forms, make the decision about whether the patient's care is medically necessary."[150] Instead, a plaintiff must show that the state actor and the private party acted jointly—for example, by carrying out "a deliberate, previously agreed upon plan" or engaging in activity constituting "a conspiracy or meeting of the minds."[151] Examples such as *Brentwood Academy* and *Blum* attempt to define the contours of the joint participation test in terms of specific human and organizational activities.[152] If the government were merely involved through standard setting but not active decisionmaking, no joint participation exists. However, as the *Brentwood* Court acknowledges, "What is fairly attributable [to the state] is a matter of normative judgment, and the criteria lack rigid simplicity."[153]

In *Cahoo*, the court found that two software companies were so entangled in allegedly unconstitutional conduct that they could potentially be found liable as state actors.[154] As to the first company, CSG

[147] See Lugar v. Edmondson Oil Co., 457 U.S. 922, 941–42 (1982).

[148] 531 U.S. 288, 298–302 (2001).

[149] Sybalski v. Indep. Grp. Home Living Program, Inc., 546 F.3d 255, 258–59 (2d Cir. 2008) (quoting Blum v. Yaretsky, 457 U.S. 991, 1006 (1982)).

[150] Id.

[151] Dahlberg v. Becker, 748 F.2d 85, 93 (2d Cir. 1984); see also West v. Atkins, 487 U.S. 42, 51 (1988) ("The Manual governing prison health care in North Carolina's institutions, which Doctor Atkins was required to observe, declares: 'The provision of health care is a joint effort of correctional administrators and health care providers, and can be achieved only through mutual trust and cooperation.'"); Wilcher v. City of Akron, 498 F.3d 516, 520 (6th Cir. 2007) (holding that the cable provider did not work "hand-in-glove" with the government to enact the challenged rules and regulations).

[152] See *Brentwood*, 531 U.S. at 295–97; *Blum*, 457 U.S. at 1006.

[153] *Brentwood*, 531 U.S. at 295.

[154] See *Cahoo I*, 322 F. Supp. 3d 772, 793 (E.D. Mich. 2018), aff'd in part and rev'd in part, 912 F.3d 887 (6th Cir. 2019). A third technology vendor, FAST, did not contest that it was a state actor for purposes of the lawsuit. See id. at 791.

Government Solutions, the court cited allegations that CSG had directly participated in the administration of unemployment benefits, "a power traditionally exclusively reserved to the State."[155] It also found that the contract between the State of Michigan and CSG had delegated "managerial authority" over the development of MiDAS to CSG as the state's application development and implementation vendor responsible for "the timely delivery of quality information technology services for all stakeholders of the [MiDAS] project."[156] The court also cited the allegations that CSG was responsible for "utilizing and mentoring" state employees on the project.[157] Thus, the court found that these allegations supported a theory that CSG, "acting in concert with the State, was 'entwined' with the UIA in administering and maintaining the robo-fraud-adjudication system that deprived the plaintiffs of their constitutional rights."[158]

As to SAS, the court found that plaintiffs had successfully pled a § 1983 claim because they had alleged that SAS, acting under the color of state law, "designed, created, implemented, maintained, configured and controlled" the EFDS, which UIA used "to make unemployment insurance fraud determinations."[159] According to the contract between the state and SAS, SAS agreed that the EFDS would "utilize[] data from the Department of Technology, Management, and Budget (DTMB)'s Data Warehouse in the development of UIA Benefit and Tax fraud detection analysis, and the results of that analysis would be integrated with MiDAS."[160] Under the terms of the contract, SAS's responsibilities included "requirements definition, functional design, configuration, testing, implementation, warranty, and maintenance."[161]

SAS responded by claiming "that it was merely an independent contractor that provided software to the State."[162] But the court found that the contract provided for more.[163] First, the court agreed with the plaintiffs' allegation that SAS did more than provide the system; it "implemented and maintained" the system.[164] Second, the court found, based on the contract, that SAS was entwined with the state because it played "a non-negligible role in the automated system."[165] For example, the contract required SAS to "schedule, coordinate, and perform all testing activities to validate that the product will operate in its intended environment,

[155] Id. at 793.
[156] Id.
[157] Id.
[158] Id.
[159] Id. at 793–94.
[160] Id. at 788.
[161] Id.
[162] Id. at 793–94.
[163] Id. at 794.
[164] Id.
[165] Id.

satisfies all user requirements, and is supported with complete and accurate operating documentation."[166] The court also highlighted the fact that "SAS was responsible for correcting defects discovered during testing and collaborating with the State to improve the system . . . [and] for providing EDFS performance tuning and defect repair."[167] Thus, the court held that, if the allegations were correct, SAS could plausibly be considered a state actor and be held constitutionally liable for the harm that MiDAS and EDFS had inflicted on the plaintiffs.[168]

While the facts in *Cahoo* make a particularly persuasive case for state action, one could also see the same "entwinement" theory applying in the disability benefits and public employment case studies.[169] In both sets of cases, the government agencies worked hand in hand with private software contractors to design, implement, and—at least in theory—test the AI system that was directly responsible for the constitutional harms involved. On the other hand, as noted above, in the criminal risk assessment context, it is less clear that providers of risk assessments are as "entwined" as CSG and SAS were in Michigan.[170] More facts would need to be known about the level of engagement and involvement in the creation, implementation, and testing of the systems.

III. When AI Systems Are More Like Private Prison Doctors than Nursing Home Administrators

Above we have outlined how, under each of the three theories of state action, courts *could* find private AI providers responsible for the constitutional harm they cause. In this section, we discuss if and when courts *should* do so.

A. *When the State Lacks Sufficient Accountability or Capacity to Provide Appropriate Remedies*

One instance in which the normative argument for constitutional accountability for private AI system vendors is the strongest is when state accountability is the weakest. For example, in the Arkansas disability

[166] Id.

[167] Id.

[168] Id. The court did, however, dismiss the individual employees of SAS, FAST, and CSG from the suit as the plaintiffs could not show that any of them had specifically and personally been involved in the actions that caused the harm. Id. This raises interesting questions about the role of humans within software engineering. Perhaps the court felt that holding the companies liable was enough and did not want to inflict too much constitutional liability on individuals, but this is hard to distinguish from *West*, especially because some humans must have been involved in the design, implementation, and testing that the court found sufficient to potentially hold CSG and SAS liable as state actors.

[169] See supra sections I.A.1–.2.

[170] See supra section I.A.3.

case,[171] the state relied on private contractors almost wholesale to design and implement the system that caused the constitutional harm.[172] While the plaintiffs were able to bring claims against the government to stop the ongoing deployment of the AI-driven program, the state agency lacked the capacity to address most of the specific causes of harm directly. The state had very little knowledge of how the AI software code had been written, where the mistakes were made, what data had been used to train and test it, or what means were required to mitigate the concerns raised in the case.[173] The same is true for the *Houston Federation of Teachers* case, in which not a single employee of the school district could explain, let alone remedy, the methods or outputs of the proprietary AI at the heart of the constitutional liability concerns.[174] This is strikingly similar in many ways to the prison doctors in *West*, to whom the state had "outsourced" its constitutional obligation to provide responsible healthcare.[175]

Of course, one could argue that this is often the case with government vendors and that holding the state accountable is often sufficient because the state can simply demand that the vendor provide the appropriate remedy or it can switch vendors. However, in *West*, we see that the Court was concerned not only with the specific harm to the single inmate in the case but also with the potential for the state to systematically avoid constitutional accountability by outsourcing potential liability to unaccountable private actors.[176] Holding private doctors personally liable ensures that they too have incentives to mitigate constitutional harms. Applying similar incentives to software vendors that sell AI systems to government agencies would accomplish similar goals. Moreover, while government actors can be assessed for damages when their AI systems violate individual rights, the primary remedy against state actors in such

[171] Ark. Dep't of Human Servs. v. Ledgerwood, 530 S.W.3d 336 (Ark. 2017).

[172] See Lecher, supra note 7 ("The instrument used in Arkansas was designed by InterRAI, a nonprofit coalition of health researchers from around the world.").

[173] AR Choices: Update & Public Comment, Medicaid Saves Lives (July 25, 2018), https://medicaidsaveslives.com/2018/07/25/ar-choices-update-public-comment/ [https://perma.cc/PD92-W3U9] (explaining that the RUGs algorithm has not been validated or verified in Arkansas); see also Letter from Kevin De Liban to Becky Murphy, supra note 55, at 4; Lecher, supra note 7.

[174] See Hous. Fed'n of Teachers, Local 2415 v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168, 1177 (S.D. Tex. 2017) (explaining that the third-party vendor "treats these algorithms and software as trade secrets, refusing to divulge them to either [the Houston Independent School District] or the teachers themselves").

[175] See Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1929 n.1 (2019) ("[A] private entity may, under certain circumstances, be deemed a state actor when the government has outsourced one of its constitutional obligations to a private entity." (citing West v. Atkins, 487 U.S. 42, 56 (1988))).

[176] See *West*, 487 U.S. at 55–58 ("Contracting out prison medical care does not relieve the State of its constitutional duty to provide adequate medical treatment to those in its custody.").

contexts is injunctive relief.[177] Unless vendors are subject to the court's jurisdiction, the court cannot assert any real oversight or impose any specific injunctive relief on that party,[178] even if it is in the best position to fix errors in how the AI performed.

B. *When AI Providers Are Underregulated*

Another normative argument in favor of finding state action applies when AI vendors are underregulated. Currently, regulatory approaches to AI are under discussion, but almost no jurisdiction has enacted rigorous regulatory approaches to ensure accountability, especially for constitutional concerns.[179] When such private accountability gaps exist, state action remedies make more normative sense. Were these gaps to be filled to allow regulators or harmed plaintiffs to sue private actors separately under alternative laws, there would be less justification for designating AI vendors with state actor status. While this was never part of the explicit holdings in previous state action cases, one can see hints of this in several of the key decisions.[180]

Again, *Cahoo* provides an excellent example. There, in the same decision upholding federal civil rights claims against CSG and SAS as state actors, the court also simultaneously dismissed all state tort claims against them. The court concluded that under Michigan's product liability, negligence, and civil conspiracy laws, neither company could potentially be held liable for its actions.[181] Therefore, the only viable claims of relief were procedural due process, equal protection, and freedom from unreasonable seizure of property.[182]

C. *When Trade Secrecy or Third-Party Technical Information Is at the Heart of the Constitutional Liability Question*

A third situation in which normative values argue in favor of state action is where trade secrecy or third-party information is at the heart of the constitutional liability question. For example, in the *Houston Federation of Teachers* case, none of the school district employees could provide any answers to the core substantive questions concerning

[177] See Corr. Servs. Corp. v. Malesko, 534 U.S. 61, 74 (2001) ("[I]njunctive relief has long been recognized as the proper means for preventing entities from acting unconstitutionally.").

[178] See Jurisdiction, Black's Law Dictionary (10th ed. 2018).

[179] See AI Now 2018 Report, supra note 10, at 39–40.

[180] But see Metzger, supra note 96, at 1425–26 (arguing that the weakness of the state action doctrine allows government actors to control private contractors without being limited by traditional constitutional constraints).

[181] *Cahoo I*, 322 F. Supp. 3d 772, 809–13 (E.D. Mich. 2018), aff'd in part and rev'd in part, 912 F.3d 887 (6th Cir. 2019).

[182] Id. at 813.

constitutional liability in the case.[183] Instead, all of those answers were within the technical and legal power of the vendor.[184] In such cases, considering the vendor a state actor would allow courts access to the necessary information to decide cases while also directly addressing vendor trade secrecy concerns. As parties, technology companies litigate their technologies every day in courts. Allowing those who have been constitutionally harmed to sue the vendors directly would allow plaintiffs and courts to access all relevant information about the AI system, its function, and the role of the vendor in the alleged constitutional violation. Vendors, of course, would have all rights to object or limit discovery under standard civil procedure provisions, including invocation of protective orders. Moreover, many courts have developed specific provisions to narrow and vet claims of trade secrecy.[185]

CONCLUSION

The state action doctrine should be considered as a potential pathway to providing greater accountability for the government use of AI systems. As Professor Gillian Metzger argues, “State action doctrine remains the primary tool courts use to ensure that private actors do not wield government power outside of constitutional constraints.”[186] As discussions of AI regulation move forward, the state action doctrine should form part of the landscape of the reasonable and appropriate regime that is ultimately devised. In particular, as AI systems rely more on deep learning, potentially becoming more autonomous and inscrutable, the accountability gap for constitutional violations threatens to become broader and deeper. This may result in both state and private human employees having less knowledge or direct involvement in the specific decisions that cause harm. For example, a new proposed rule from the U.S. Department of Housing and Urban Development creates a complete defense to a prima facie case of housing discrimination when the defendant uses an industry-standard algorithmic model to make its housing decisions.[187] This rule, if adopted, would encourage many actors in the housing industry to use AI systems, knowing that they could avoid liability

[183] See Hous. Fed’n of Teachers, Local 2415 v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168, 1177 (S.D. Tex. 2017).

[184] See id.

[185] See Cal. Civ. Proc. Code § 2019.210 (2019) (requiring trade-secrecy plaintiffs to identify the trade secret with reasonable particularity before allowing any discovery related to the trade secret to proceed).

[186] Metzger, supra note 96, at 1410.

[187] See Michelle Aronowitz & Olatunde Johnson, The Trump Administration’s Assault on Fair Housing, Take Care (Aug. 19, 2019), https://takecareblog.com/blog/the-trump-administration-s-assault-on-fair-housing [https://perma.cc/23J7WVWW]; see also Andrew D. Selbst, A New HUD Rule Would Effectively Encourage Discrimination by Algorithm, Slate (Aug. 19, 2019), https://slate.com/technology/2019/08/hud-disparate-impact-discrimination-algorithm.html [https://perma.cc/X8NE-Q2Z6].

by blaming the AI itself, even if there was overwhelming evidence that they knew the use of the system would have a disparate discriminatory impact.[188]

No doubt there will be many attempts, such as the proposed HUD rule, to allow AI systems to be used as accountability-avoidance mechanisms when companies cause constitutional violations. This is why the state action doctrine must remain a powerful and flexible common law approach for courts to use to redress this gap as it widens. This will be particularly necessary if legislation or agency regulation is slow to materialize or inadequate for the complex task that AI will present in the coming years.

[188] Cf. Logiodice v. Trs. of MCI, 296 F.3d 22, 27 (1st Cir. 2002) (holding private entity MCI was not a state actor even though it ran part of the Maine public school system because the alleged constitutional violation—illegally disciplining a student—was carried out entirely and internally by MCI and had no joint participation elements). But see Patrick v. Success Acad., 354 F. Supp. 3d 185, 209 n.24 (2018) (noting that private operators of charter schools are state actors, despite extensive internalization of decisionmaking).

DISRUPTIVE INCUMBENTS: PLATFORM COMPETITION IN AN AGE OF MACHINE LEARNING

C. Scott Hemphill *

Recent advances in machine learning have reinforced the competitive position of leading online platforms. This Essay identifies two important sources of platform rivalry and proposes ways to maximize their competitive potential under existing antitrust law. A nascent competitor is a threatening new entrant that, in time, might become a full-fledged platform rival. A platform's acquisition of a nascent competitor should be prohibited as an unlawful acquisition or maintenance of monopoly. A disruptive incumbent is an established firm—often another platform—that introduces fresh competition in an adjacent market. Antitrust enforcers should take a more cautious approach, on the margin, when evaluating actions taken by a disruptive incumbent to compete with an entrenched platform.

INTRODUCTION

The leading online platforms—Google in search, Facebook in social network services, and Amazon in e-commerce—benefit from economies of scale and access to user data that are difficult for rivals to replicate. These barriers are reinforced by advances in machine learning, a set of artificial intelligence (AI) techniques[1] that use models to "learn" desired behavior from "examples rather than instructions."[2] This Essay considers how competition might be enhanced, notwithstanding these advantages, under existing antitrust law.[3]

[*] Moses H. Grossman Professor of Law, New York University School of Law. I thank John Asker, Adam Cox, Harry First, Jacob Gersen, Jeannie Gersen, Bert Huang, Avery Katz, Benedict Kingsbury, Bhaven Sampat, Tim Wu, and audiences at Columbia, ETH Zurich, and NYU, for helpful comments. Tim Keegan, Ryan Knox, Ina Kosova, Alison Perry, and David Stein provided outstanding research assistance.

[1] As used here, artificial intelligence refers to technologies that mimic or resemble some aspect of human intelligence. In some contexts, the AI label can be misleading, given that the task at issue—for example, online search—was automated to begin with, and the deployment of improved software does not entail any direct replacement of labor. See Timothy Bresnahan, Artificial Intelligence Technologies and Aggregate Growth Prospects 2 (May 2019) (unpublished manuscript) (on file with the *Columbia Law Review*) (discussing this issue).

[2] Machine Learning, IBM Design for AI, https://ai-design.eu-de.mybluemix.net/design/ai/basics/ml [https://perma.cc/T5NK-TCU7] (last visited Sept. 30, 2019). See generally A.L. Samuel, Some Studies in Machine Learning Using the Game of Checkers, 3 IBM J. Res. & Dev. 211 (1959) (coining the term "machine learning"). For further discussion, see infra section I.A.

[3] Machine learning also challenges antitrust policy by facilitating collusion and price discrimination. For a discussion, see generally Ariel Ezrachi & Maurice Stucke,

Two sources of platform competition are particularly important. A *nascent competitor* is a threatening new entrant that, in time, might become a full-fledged platform rival. For example, Instagram posed an important threat to Facebook shortly after Instagram's launch in 2010. A *disruptive incumbent* is an established firm, often another platform, that introduces fresh competition in an adjacent platform market. For example, Microsoft's Bing search platform competes with Google's. In turn, Google vies with Amazon for so-called shopping starts—that is, to be the starting place for online shoppers.

Antitrust law protects nascent competitors as a source of platform entry.[4] This Essay argues that the Sherman Act prohibits the acquisition of a nascent competitor as a form of unlawful monopolization.[5] Monopolization, a branch of antitrust law typically concerned with exclusionary conduct, also reaches acquisitions and other cooperative behavior. The law extends to both newly announced mergers and other transactions, such as Facebook's acquisition of Instagram in 2012, that have been consummated. Some transactions also violate Section 7 of the Clayton Act, the statute ordinarily relied upon to prohibit unlawful mergers.[6] The Sherman Act approach, however, is a better fit for the evaluation of some acquisitions, due in part to judicial recognition that the target need not operate in the same antitrust market as the acquirer.[7]

Disruptive incumbents are a second, and underappreciated, source of platform competition.[8] A disruptive incumbent is well positioned to compete with a dominant platform in an adjacent market. Such firms can deploy a variety of large-firm advantages without fear of cannibalizing their home market. Thus, disruptive incumbents sidestep a longstanding debate, associated with economists Joseph Schumpeter and Kenneth Arrow, about whether monopoly or competition best promotes innovation.[9] This Essay suggests that antitrust enforcers should consider a lighter touch toward enforcement, on the margin, if such a firm is "punching up" to compete with a platform—think of Google presenting shopping search results in a particular (by assumption, legally contestable) manner to better compete with Amazon.

As challengers, neither nascent competitors nor disruptive incumbents are sure things. Instagram, absent the acquisition, might have failed to compete with Facebook. Google might ultimately lose its battle with Amazon for shopping starts, even if antitrust enforcers leave this aspect of its conduct alone. Given the potentially large benefits of

Virtual Competition: The Promise and Perils of the Algorithm-Driven Economy (2016). These developments are beyond the scope of this Essay.

[4] See infra Part II.

[5] See 15 U.S.C. § 2 (2012).

[6] See id. § 18.

[7] See infra section II.B.

[8] See infra Part III.

[9] See infra section III.A.

successful competition from a disruptive incumbent, a modest probability of success may sometimes justify a lighter touch, provided that the negative collateral consequences—and, to be clear, there may be some—are not too large.

This Essay proceeds in three parts. Part I spells out several barriers to platform entry, emphasizing the role of machine learning, and the benefits of increased competition. Part II makes the case that a platform's acquisition of a nascent competitor may constitute unlawful monopolization. Part III explains the role of disruptive incumbents and their relevance to the Arrow–Schumpeter debate, suggesting conditions under which their conduct might merit a lighter touch from antitrust enforcers.

I. Promoting Platform Entry

A. *Machine Learning as a Barrier to Entry*

Leading platforms make money by matching users with advertisers and products.[10] Google and Facebook display ads alongside other content, such as Google's unpaid "organic" search results and Facebook's news feed.[11] The platform is typically paid when a user clicks on the ad.[12] Amazon matches users with recommendations about products available for purchase and makes money from successfully completed purchases.[13]

The matching process is driven in part by algorithms that predict the likelihood that a user will click on an ad or buy a product. Google runs a complex auction among advertisers that vie for placement on the search engine results page in response to a user search.[14] For each proposed ad, Google calculates an "Ad Rank," which incorporates a prediction about

[10] Ajay Agrawal, Joshua Gans & Avi Goldfarb, Prediction Machines: The Simple Economics of Artificial Intelligence 83 (2018); Bresnahan, supra note 1, at 6.

[11] Google and (to a lesser degree) Facebook also place ads on independent websites and mobile apps. See Display Campaigns, Google Ads, https://ads.google.com/home/campaigns/display-ads [https://perma.cc/92EQ-BV5S] (last visited Sept. 30, 2019) (describing display ads placed by Google on third-party websites); Google, How AdSense Works, AdSense Help, https://support.google.com/adsense/answer/6242051 [https://perma.cc/W2AB-LYQW] (last visited Sept. 30, 2019) (describing search ads placed by Google on third-party websites); see also Audience Network by Facebook, Facebook, https://www.facebook.com/audiencenetwork/products [https://perma.cc/F8M7-TK3X] (last visited Sept. 30, 2019) (describing a Facebook ad product delivered to mobile apps and mobile websites).

[12] Other payment models similarly rely upon a measurable user action, such as making a purchase. See Magdalena Rzemieniak, Measuring the Effectiveness of Online Advertising Campaigns in the Aspect of E-Entrepreneurship, 65 Procedia Computer Sci. 980, 981–83 (2015) (reviewing various payment models).

[13] Amazon also makes money from ads and from marketing expenditures by sellers. Amazon.com, Inc., Annual Report (Form 10-K) 42 (Jan. 31, 2019) [hereinafter Amazon 2018 Annual Report] (describing revenue from pay-per-click and pay-per-impression ads).

[14] About Ad Position and Ad Rank, Google Ads Help, https://support.google.com/google-ads/answer/1722122 [https://perma.cc/J5QK-YTCJ] (last visited Sept. 30, 2019).

the probability that a user will click on the ad.[15] Facebook's ad system similarly relies on a prediction of the clickthrough rate.[16] Amazon's system relies upon an analogous prediction about which products are most likely to interest a user.[17]

Machine learning improves the predictions of the matching algorithms. The learning occurs inductively—that is, bottom up—via automated evaluation of the examples.[18] For example, Google uses machine learning—incorporating what Google knows about the user, the search term, and the ad—to help predict the clickthrough rate for a specific user as to a specific ad.[19] Amazon likewise uses machine learning to improve its product recommendations.[20] The leading platforms have been relying on machine learning to improve their predictions for some years, which is hardly surprising, given that this tool is well suited to concrete metrics such as the clickthrough rate or purchase rate. Moreover, erroneous predictions are not very costly. If the suggestion is inapt, that simply means that a user does not click or buy.[21]

The leading platforms have avidly pursued investments in machine learning. Alphabet (Google's parent company) places machine learning front and center in its 2018 annual report,[22] and the head of AI reports directly to Google's CEO.[23] According to its CEO, Google is "applying

[15] All else equal, a more frequently clicked ad is more valuable to Google. Id. The organic results are generated by other complex algorithms, which were originally based on PageRank but now factor in hundreds of other signals. How Search Algorithms Work, Google Search, https://www.google.com/search/howsearchworks/algorithms [https://perma.cc/R54W-CFAR] (last visited Sept. 30, 2019).

[16] Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers & Joaquin Quiñonero Candela, Practical Lessons from Predicting Clicks on Ads at Facebook, Facebook Research (Aug. 24, 2014), https://research.fb.com/wp-content/uploads/2016/11/practical-lessons-from-predicting-clicks-on-ads-at-facebook.pdf [https://perma.cc/6R9W-GCC7].

[17] Brent Smith & Greg Linden, Two Decades of Recommender Systems at Amazon.com, IEEE Internet Computing, May–June 2017, at 12, 12.

[18] See Samuel, supra note 2, at 211 (defining "machine learning" by reference to "[p]rogramming computers to learn from experience" rather than specifying a "solution in minute and exact detail").

[19] Matt Lawson, Grow Your Business Faster with Machine Learning: Part I, Google Ads Blog (Jan. 8, 2018), https://www.blog.google/products/ads/adwords-machine-learning-part-1 [https://perma.cc/79ZF-MAG2].

[20] Jeffrey P. Bezos, 2016 Letter to Shareholders, Amazon Blog: Day One (Apr. 17, 2017), https://blog.aboutamazon.com/company-news/2016-letter-to-shareholders [https://perma.cc/C6GG-G9RX] (explaining that machine learning "drives" product recommendation algorithms, among others).

[21] Bresnahan, supra note 1, at 11. Contrast this with the high cost of a failure to automatically filter inappropriate content out of a Facebook news feed. Id. at 17–18.

[22] Alphabet Inc., Annual Report (Form 10-K) 3 (Feb. 4, 2019) ("Across the company, machine learning and [AI] are increasingly driving many of our latest innovations.").

[23] See Google: Org Chart, The Org, https://theorg.com/org/google [https://perma.cc/Z6JU-HAZN] (last visited Sept. 30, 2019) (showing Google's AI Lead among the CEO's direct reports).

machine learning and AI . . . across every one of [its] products" as part of an "AI-first approach."[24] Amazon uses machine learning to forecast demand and place fulfilment centers, among other tasks.[25] Salaries for scarce technical talent have skyrocketed.[26] Some of these investments and advances pertain to so-called "deep learning," a set of machine learning techniques that make domain expertise less important.[27] Facebook's technical infrastructure is reportedly "entirely built around" deep learning.[28] At Google, the introduction of deep learning rapidly doubled its computational load.[29]

Advances in machine learning reinforce the strong position already enjoyed by the leading platforms. Making an improvement by this method has a high fixed cost and low marginal cost, a combination that tends to favor large firms that can spread the fixed cost over a large number of units. A firm with a large existing base of users is particularly well

[24] Google Developers, Google I/O Keynote (Google I/O '17), YouTube (May 17, 2017), https://www.youtube.com/watch?v=Y2VF8tmLFHw (on file with the *Columbia Law Review*) ("[I]n an AI-first world, we are rethinking all our products and applying machine learning and AI to solve user problems. And we are doing this across every one of our products."); see also Ross Kelly, Forget Putting Mobile First, It's All About AI These Days: Google CEO, Chief Executive (May 18, 2017), https://chiefexecutive.net/forget-putting-mobile-first-ai-days-google-ceo [https://perma.cc/7MXF-6LLD] (reporting Google CEO's remarks).

[25] See Machine Learning Center of Excellence, Amazon Jobs, https://www.amazon.jobs/en/teams/machine-learning [https://perma.cc/AH5X-36XJ] (last visited Sept. 30, 2019).

[26] Cade Metz, Tech Giants Are Paying Huge Salaries for Scarce A.I. Talent, N.Y. Times (Oct. 22, 2017), https://www.nytimes.com/2017/10/22/technology/artificial-intelligence-experts-salaries.html [https://perma.cc/J8LG-8VMF] ("Salaries are spiraling so fast that some joke the tech industry needs a National Football League-style salary cap on A.I. specialists.").

[27] Deep learning is a class of machine learning techniques that process examples with relatively little domain-specific instruction from the implementer. For example, perspective presents a serious difficulty in image-labeling tasks. A traditional machine learning approach to image labeling might include a detailed model of perspective. By contrast, a deep learning approach would dispense with the need for such a model in its initial configuration. Instead, the software arrives at its own method for overcoming the difficulties. For a seminal paper illustrating this approach, see Alex Krizhevsky, Ilya Sutskever & Geoffrey Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Comm. ACM, June 2017, at 84, 84–85 (describing an image recognition system without reliance upon image-processing-specific logic); see also Hal Varian, Artificial Intelligence, and Industrial Organization, *in* The Economics of Artificial Intelligence: An Agenda 399, 399–400 (Ajay Agrawal, Joshua Gans & Avi Goldfarb eds., 2019) (providing a brief explanation of the deep learning approach).

[28] Rachel Metz, Facebook's Top AI Scientist Says It's "Dust" Without Artificial Intelligence, CNN Bus., https://www.cnn.com/2018/12/05/tech/ai-facebook-lecun/index.html [https://perma.cc/FG78-HMU8] (last updated Dec. 5, 2018) ("The technology is included in everything from the posts and translations you see in your news feed to advertisements.").

[29] Jeff Dean, David Patterson & Cliff Young, A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution, IEEE Micro, Mar./Apr. 2018, at 21, 22 (describing a doubling in "computation demands" due to increased use of "deep neural networks," and the development of custom hardware to handle these demands).

positioned to profit from—and hence, incentivized to pursue—any incremental benefit. Even a small improvement can make a big difference to the bottom line. The same argument applies to custom hardware to support machine learning, which Google and others have invested in to provide greater processing power at a given cost.[30]

Machine learning advances also reinforce the importance of access to data.[31] A larger stock of searches and observed outcomes—for example, whether the user clicked—generates data needed to train and improve the prediction of the algorithm.[32] The importance of scale is heightened by the high variability of user data.[33] With too few queries, it is difficult to train the algorithm to match queries effectively.[34] This advantage is subject to the limiting principle that eventually there are decreasing returns to scale.[35]

A particular type of data, user history, is important in some applications. Recommendations and ads reflect inferences based on a user's previous purchases and searches.[36] Moreover, the past behavior of a large set of users provides a helpful starting point for predicting the behavior of an individual user. For example, Amazon suggests that a purchaser of

[30] See Norman P. Jouppi, Cliff Young, Mishant Patil & David Patterson, Motivation for and Evaluation of the First Tensor Processing Unit, IEEE Micro, May/June 2018, at 10, 14 tbl.2, 16 tbl.4 (reporting much higher performance and lower power usage for custom hardware, compared to traditional hardware); see also Kalin Ovtcharov, Olatunji Ruwase, Joo-Young Kim, Jeremy Flowers, Karin Strauss & Eric S. Chung, Microsoft, Toward Accelerating Deep Learning at Scale Using Specialized Hardware in the Datacenter 7 (2015) (on file with the *Columbia Law Review*) (noting that servers with chips that support the use of machine learning result in "low overhead in power and cost per server").

[31] Judith Chevalier, Comment on "Artificial Intelligence, Economics, and Industrial Organization," *in* The Economics of Artificial Intelligence, supra note 27, at 419, 419 (emphasizing lack of data access as a barrier to entry).

[32] He et al., supra note 16, fig.10 (showing lower quality of prediction when using only one percent of Facebook training data).

[33] Google collects a set of data points, or "features" of a user, to help predict the clickthrough rate. H. Brendan McMahan, Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos & Jeremy Kubica, Ad Click Prediction: A View from the Trenches, Google Research § 4 (2013), https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/41159.pdf [https://perma.cc/L5US-826P]. Even at Google's massive scale, about half of the features collected are completely unique and therefore lack predictive power. Id.

[34] See He et al., supra note 16, fig.10 (describing increased accuracy accomplished with a larger dataset).

[35] See Varian, supra note 27, at 406 (applying to machine learning the "general principle" that "data typically exhibits decreasing returns to scale like any other factor of production").

[36] See, e.g., Greg Linden, Brent Smith & Jeremy York, Amazon.com Recommendations: Item-to-Item Collaborative Filtering, IEEE Internet Computing, Jan.–Feb. 2003, at 76, 78 ("Given the user's purchased and rated items, the algorithm constructs a search query to find other popular items by the same author, artist, or director, or with similar keywords or subjects.").

Prediction Machines: The Simple Economics of Artificial Intelligence[37] might also want to buy *Applied Artificial Intelligence*[38] because past customers have done so.[39] The importance of user history varies by application, and more recent user data often have an outsized importance.[40] Still, user history is a resource that an entrant cannot easily replicate. Absent data portability, this information is difficult to acquire even at a high price.[41]

Machine learning may reduce the need to retain historical data, or so much of it. For certain models, once past results are incorporated into the model, no further use of the historical data is made when making further predictions.[42] As a fanciful example, if a user buys only books about artificial intelligence, making use of that insight for prediction does not require referring back to the full list of past purchases. Historical data are still useful to train a new model, posing a downside to simply discarding the data.

Reduced reliance on historical data, including personally identifiable information, would ease privacy concerns stemming from the retention of such information.[43] This technical possibility, however, does not lower the barrier to entry for other firms, which still lack access to the historical data now incorporated into the incumbent's algorithm. By undermining the effectiveness of access and portability proposals, which rely on the transfer of user data as a way to jumpstart competition, certain barriers to entry may actually increase.[44]

[37] Agrawal et al., supra note 10.

[38] Mariya Yao, Marlene Jia & Adelyn Zhou, Applied Artificial Intelligence: A Handbook for Business Leaders (2018).

[39] Prediction Machines: The Simple Economics of Artificial Intelligence, Amazon, https://www.amazon.com/Prediction-Machines-Economics-Artificial-Intelligence/dp/1633695670 [https://perma.cc/FFT5-EVTW] (last visited Sept. 30, 2019) (showing, on the product page, that customers often buy *Applied Artificial Intelligence* after viewing *Prediction Machines*).

[40] See Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, James Law, Kevin Lee, Jason Lu, Pieter Noordhuis, Misha Smelyanskiy, Liang Xiong & Xiaodong Wang, Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective, Facebook Research § IV.B (2018), https://research.fb.com/wp-content/uploads/2017/12/hpca-2018-facebook.pdf [https://perma.cc/8MQY-LN3K] (describing problems arising from the use of "stale models"); He et al., supra note 16, § 3.2 (reporting that "[p]rediction accuracy clearly degrades" with the use of older training data).

[41] See Daniel L. Rubinfeld & Michal S. Gal, Access Barriers to Big Data, 59 Ariz. L. Rev. 339, 351 (2017) ("Another barrier may be temporal, relating to the point in time that the competitor started gathering data. To illustrate, a collection of aerial maps before a natural disaster cannot be replicated once the disaster occurs.").

[42] See, e.g., McMahan et al., supra note 33, § 3 (describing a class of models in which "each training example only needs to be considered once").

[43] See Nat'l Inst. of Standards & Tech., U.S. Dep't of Commerce, NIST Special Pub. 800-122, Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) 4-3 (2010) (outlining the privacy risks of using personally identifiable information).

[44] One possible solution is the use of so-called federated learning techniques, in which a user's data are kept locally on the user's machine without the algorithm provider

Improving the matching algorithm is an important application of machine learning, but there are many others. For example, Google relies on machine learning to rank queries in its organic search algorithm,[45] translate webpages,[46] and suggest user responses in email.[47] Deep learning techniques have greatly improved automatic speech recognition,[48] a key input for digital assistants and other voice-based user interfaces.[49] In various ways, these developments, too, may tend to reinforce the position of the leading platforms.

Not all machine learning developments strengthen the barriers to entry. Advances in speech recognition have been achieved using public datasets and do not necessarily require data at a massive scale.[50] Proprietary datasets are sometimes released to the public, thereby enabling innovation on a decentralized basis.[51] Some striking developments have been made using small teams. For example, in 2012, a group at the University of Toronto achieved a major breakthrough in image labeling, using deep learning techniques that are now at the core of the machine

having direct access to the data. See Brendan McMahan & Daniel Ramage, Federated Learning: Collaborative Machine Learning Without Centralized Training Data, Google AI Blog (Apr. 6, 2017), https://ai.googleblog.com/2017/04/federated-learning-collaborative.html [https://perma.cc/GU2Z-QZJH]. Aside from protecting privacy, federated learning might have the further benefit of increasing the effectiveness of data portability.

[45] See Jack Clark, Google Turning Its Lucrative Web Search Over to AI Machines, Bloomberg (Oct. 26, 2015), https://www.bloomberg.com/news/articles/2015-10-26/google-turning-its-lucrative-web-search-over-to-ai-machines [https://perma.cc/9362-SKTY].

[46] See Mike Schuster, Melvin Johnson & Nikhil Thorat, Zero-Shot Translation with Google's Multilingual Neural Machine Translation System, Google AI Blog (Nov. 22, 2016), https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html [https://perma.cc/6EDU-A9SS].

[47] See Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, László Lukács, Marina Ganea, Peter Young & Vivek Ramavajjala, Google Inc., Smart Reply: Automated Response Suggestion for Email, SIGKDD § 1 (2016), https://www.kdd.org/kdd2016/papers/files/Paper_1069.pdf [https://perma.cc/B2NP-339K] (describing the architecture of Gmail's Smart Reply system).

[48] See, e.g., W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu & G. Zweig, Microsoft Research, Achieving Human Parity in Conversational Speech Recognition, arXiv § 9 (2017), https://arxiv.org/pdf/1610.05256.pdf [https://perma.cc/WV95-MGM3] ("We find that the machine errors are substantially the same as human ones . . . .").

[49] See Bresnahan, supra note 1, at 25–29 (characterizing this development). Digital assistants are offered (as of September 2019) by Alibaba (Genie), Amazon (Alexa), Apple (Siri), Google (Google Assistant), and Microsoft (Cortana).

[50] See, e.g., Xiong et al., supra note 48, § 1 (demonstrating machine performance that is on par with human speech recognition performance using a public dataset).

[51] See, e.g., Luc Vincent, Unlocking Access to Self-Driving Research: The Lyft Level 5 Dataset and Competition, Medium (July 23, 2019), https://medium.com/lyftlevel5/unlocking-access-to-self-driving-research-the-lyft-level-5-dataset-and-competition-d487c27b1b6c [https://perma.cc/2F64-Z49P] (describing the release of one such dataset in order to facilitate autonomous driving research).

learning efforts of the leading platforms.[52] And although many machine language advances and insights by the incumbents are held as trade secrets, there are important knowledge spillovers that benefit everyone working in these fields.

Moreover, incumbents have placed certain advances in machine learning at the disposal of other firms. For example, a major revenue stream for Amazon is Amazon Web Services (AWS), which sells storage and computing power to other businesses.[53] AWS has enabled a great deal of innovation by other firms. For example, the leading file storage firm Dropbox, for much of its existence, relied on AWS to do the actual file storage.[54] AWS also offers pretrained machine learning models on its platform.[55] Google, meanwhile, sells time on its custom hardware[56] and has released a free suite of tools to facilitate the development of new machine learning systems.[57]

Considered as a whole, advances in machine learning tend to reinforce the market position of the leading platforms. There is reason to agree with the *Economist*'s assessment, emphasizing various advantages of the incumbents: "It seems likely that the incumbent tech groups will capture many of AI's gains, given their wealth of data, computing power, smart algorithms and human talent, not to mention a head start on investing."[58]

B. *Encouraging Entry*

Fostering competition against the leading platforms is socially desirable for several reasons. First, competition encourages lower prices and higher quality on both sides of the platform, including lower prices to advertisers and greater privacy protection for users.

[52] See From Not Working to Neural Networking, Economist (June 23, 2016), https://www.economist.com/special-report/2016/06/23/from-not-working-to-neural-networking [https://perma.cc/V2Tx-MUN3].

[53] See infra note 123 and accompanying text.

[54] See Akhil Gupta, Scaling to Exabytes and Beyond, Dropbox (Mar. 14, 2016), https://blogs.dropbox.com/tech/2016/03/magic-pocket-infrastructure [https://perma.cc/SYM9-R846].

[55] Machine Learning on AWS: Putting Machine Learning in the Hands of Every Developer, Amazon, https://aws.amazon.com/machine-learning [https://perma.cc/363Y-UJGZ] (last visited Sept. 30, 2019).

[56] Cloud TPU, Google Cloud, https://cloud.google.com/tpu [https://perma.cc/VX7P-V7PM] (last visited Sept. 30, 2019).

[57] See An End-to-End Open Source Machine Learning Platform, TensorFlow, https://www.tensorflow.org [https://perma.cc/8378-3ZXT] (last visited Sept. 30, 2019) (describing TensorFlow, an "end-to-end open source machine learning platform").

[58] Google Leads in the Race to Dominate Artificial Intelligence, Economist (Dec. 7, 2017), https://www.economist.com/business/2017/12/07/google-leads-in-the-race-to-dominate-artificial-intelligence [https://perma.cc/AAB5-GYFC].

Second, competition spurs innovation.[59] Incumbents are reluctant to cannibalize their existing business. Just think of a mobile telephone service provider considering whether to offer a communications application that it is unable to monetize or control, or a book publisher considering whether to disintermediate itself through self-publishing. As Arrow showed, an incumbent's incentive to innovate is lessened by a "replacement effect," so named because the resulting innovation replaces existing profitable sales.[60] The replacement effect suggests that innovations are more likely to come from an outsider with no existing sales to replace. An incumbent also may be hamstrung by its own earlier success.[61]

Third, competition curbs an incumbent's ability to engage in anticompetitive, self-entrenching conduct. A powerful incumbent may possess the incentive and ability, unless restrained, to starve a rival of access to inputs or customers.[62] This possibility motivates the European Commission's scrutiny of Google and Amazon search results that allegedly favor the platform's own offerings over those of third parties.[63]

These anticipated benefits have prompted commentators to offer a wide range of new policies to increase competition. The proposals range from clear labeling of search results and recommendations, mandated

[59] A counterargument, that incumbency confers a strong incentive and capacity to innovate, is considered infra Part III.

[60] See Kenneth Arrow, Economic Welfare and the Allocation of Resources for Invention, *in* The Rate and Direction of Inventive Activity: Economic and Social Factors 609, 619–22 (Richard R. Nelson ed., 1962) ("The preinvention monopoly power acts as a strong disincentive to further innovation."); see also Jean Tirole, The Theory of Industrial Organization 392 (1988) (naming Arrow's argument the "replacement effect"). A similar effect occurs when the switchover to a new product requires a temporary halt to profitable sales. See Thomas J. Holmes, David K. Levine & James A. Schmitz Jr., Monopoly and the Incentive to Innovate When Adoption Involves Switchover Disruptions, Am. Econ. J., Aug. 2012, at 1, 11–12.

[61] Clayton M. Christensen, The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail, at xv (1997) (advancing the thesis that a successful firm's focus on its current profitable customers may cause it to neglect other opportunities).

[62] See C. Scott Hemphill & Tim Wu, Parallel Exclusion, 122 Yale L.J. 1182, 1200–09 (2013) (discussing mechanisms through which an incumbent may exclude an entrant or rival). Competition does not necessarily curb exclusion, however, if the firms share an interest in exclusion. See id. at 1219–35 (describing firms' incentives and ability to engage in collective exclusion).

[63] Commission Decision of June 27, 2017 (Case AT. 39740—Google Search (Shopping)), 2018 O.J. (C 009) ¶¶ 11–14, https://ec.europa.eu/competition/antitrust/cases/dec_docs/39740/39740_14996_3.pdf [https://perma.cc/9LF4-CQ8X] [hereinafter Google Shopping Decision] (finding unlawful Google's promotion of specialized product results over third-party comparison-shopping engines and imposing a €2.4 billion fine); Press Release IP/19/4291, European Comm'n, Antitrust: Commission Opens Investigation into Possible Anti-Competitive Conduct of Amazon (July 17, 2019), https://europa.eu/rapid/press-release_IP-19-4291_en.htm [https://perma.cc/6PAX-QA5F] (raising concerns about Amazon's alleged manipulation of search results to favor its own private label products and retail operation).

data portability, and traditional utility regulation,[64] to new legislation to break up leading platforms.[65] My goal is complementary, albeit more modest—to consider how existing antitrust law can facilitate platform competition. Two sets of competitors present themselves: nascent platform rivals and disruptive incumbents. I consider these in turn.

II. Nascent Competitors

A. *Facebook's Acquisition of Instagram*

Facebook is the world's dominant social network provider, with more than 2 billion users,[66] a market capitalization in excess of $500 billion,[67] and more than $50 billion in annual revenue.[68] In 2012, Facebook reached a $1 billion deal to acquire Instagram, a photo-sharing app for mobile devices introduced in 2010.[69] The Instagram app made mobile photo sharing easy and fun. Its growth was explosive, rising from 100,000 users in October 2010[70] to 40 million in April 2012.[71]

Success in mobile was hugely important to Facebook's future prospects.[72] Facebook, however, found the move from desktop to mobile to be a difficult challenge and was slow to add a visual element to its offerings.[73] Better photo sharing was a potentially compelling reason for users

[64] See Samuel Himel & Robert Seamans, Competition Policy Int'l, Artificial Intelligence, Incentives to Innovate, and Competition Policy 4–10 (2017), https://www.competitionpolicyinternational.com/wp-content/uploads/2017/12/CPI-Himel-Seamans.pdf [https://perma.cc/82G5-L25N] (discussing a range of proposals); see also Marvin Ammori, The FTC Should Take a Broader Look at Transparency, Gigaom (June 23, 2012), https://gigaom.com/2012/06/23/the-ftc-should-take-a-broader-look-at-transparency [https://perma.cc/S7CC-N65U] (calling for more transparency and clear labeling by search engines).

[65] Elizabeth Warren, Here's How We Can Break Up Big Tech, Medium (Mar. 8, 2019), https://medium.com/@teamwarren/heres-how-we-can-break-up-big-tech-9ad9e0da324c [https://perma.cc/4MXA-DN4Z].

[66] Facebook, Inc., Quarterly Report (Form 10-Q) 25 (July 25, 2019) [hereinafter Facebook Quarterly Report] (reporting 2.4 billion monthly active users as of June 30, 2019).

[67] Facebook, Inc., Yahoo Fin., https://finance.yahoo.com/quote/FB [https://perma.cc/2XBT-2BF7] (last visited Oct. 16, 2019) (reporting market capitalization of $538 billion on Oct. 16, 2019).

[68] Facebook, Inc., Annual Report (Form 10-K) 33 (Jan. 31, 2019) (reporting $55.8 billion in revenue in 2018).

[69] Kara Swisher, The Money Shot, Vanity Fair (May 6, 2013), https://www.vanityfair.com/news/business/2013/06/kara-swisher-instagram [https://perma.cc/2GDV-26UQ].

[70] How Many Users Does Instagram Have?, Quora (Oct. 13, 2010), https://www.quora.com/How-many-users-does-Instagram-have [https://perma.cc/V33J-RGKS] (indicating 100,000 active users according to CEO Kevin Systrom).

[71] See Matt Burns, Instagram's User Count Now at 40 Million, Saw 10 Million New Users in Last 10 Days, TechCrunch (Apr. 13, 2012), https://techcrunch.com/2012/04/13/instagrams-user-count-now-at-40-million-saw-10-million-new-users-in-last-10-days [https://perma.cc/6UNW-K29K].

[72] In the second quarter of 2019, 94% of Facebook advertising revenue was from mobile. Facebook Quarterly Report, supra note 66, at 33.

[73] See Swisher, supra note 69.

to prefer Instagram to Facebook. Commentators were quick to recognize this threat. One explained that "Instagram had found and attacked Facebook's [A]chilles heel—mobile photo sharing."[74] Another noted that Instagram's advantages—"hip, elegant, fun, and 'mobile-first'"—posed a "very real threat" to Facebook.[75]

Publicly available information suggests that Facebook recognized the Instagram threat and that its acquisition may have been aimed at its elimination. A top Facebook official reportedly wrote colleagues that the purpose of the transaction was "to eliminate a potential competitor."[76] As a contemporaneous commentator explained, Facebook recognized that "for [the] first time in its life it arguably had a competitor that could not only eat its lunch, but also destroy its future prospects."[77] The FTC and other antitrust enforcers investigated but ultimately declined to challenge the acquisition.[78]

B. *Acquisitions to Acquire or Maintain a Monopoly*

Section 2 of the Sherman Act, which prohibits "monopoliz[ation],"[79] is an appropriate framework for evaluating acquisitions by an incumbent platform. The leading modern Section 2 case is *United States v. Microsoft*.[80] In 1998, the Justice Department and plaintiff states filed suit alleging that Microsoft had identified an emergent threat to its Windows operating system monopoly and had taken improper actions to neutralize it.[81]

The danger posed by the growth of the internet was articulated by CEO Bill Gates in an internal memo to his top lieutenants describing a coming "Internet Tidal Wave."[82] Netscape's browser was the centerpiece of that threat, even though Netscape's offering did not compete with

[74] Om Malik, Here Is Why Facebook Bought Instagram, Gigaom (Apr. 9, 2012), https://gigaom.com/2012/04/09/here-is-why-did-facebook-bought-instagram [https://perma.cc/Z6Z3-AL2S].

[75] Swisher, supra note 69.

[76] Josh Kosman, Facebook Boasted of Buying Instagram to Kill the Competition: Sources, N.Y. Post (Feb. 26, 2019), https://nypost.com/2019/02/26/facebook-boasted-of-buying-instagram-to-kill-the-competition-sources [https://perma.cc/E4QG-Y3S9] (reporting the contents of a document uncovered during FTC review of the transaction).

[77] Malik, supra note 74.

[78] Press Release, FTC, FTC Closes Its Investigation into Facebook's Proposed Acquisition of Instagram Photo Sharing Program (Aug. 22, 2012), https://www.ftc.gov/news-events/press-releases/2012/08/ftc-closes-its-investigation-facebooks-proposed-acquisition [https://perma.cc/J6YC-44GG].

[79] 15 U.S.C. § 2 (2012).

[80] United States v. Microsoft Corp. (*Microsoft II*), 253 F.3d 34 (D.C. Cir. 2001) (en banc) (per curiam).

[81] Id. at 47.

[82] Memorandum from Bill Gates to Exec. Staff of Microsoft Corp. 1 (May 26, 1995), https://www.justice.gov/sites/default/files/atr/legacy/2006/03/03/20.pdf [https://perma.cc/Y3Q9-CU64].

Windows.[83] Following a bench trial, the district court held that Microsoft had committed various acts, such as the suppression of browser distribution, to improperly neutralize the Netscape threat and maintain its monopoly.[84] The D.C. Circuit affirmed certain elements of the liability determination, while reversing on others.[85]

The *Microsoft* precedent illuminates why the acquisition of a nascent competitor, made to acquire or maintain a monopoly, violates Section 2. Three features of Section 2 law, in the context of a government suit seeking injunctive relief, are particularly important. First, the competitive threat posed by the target need not be fully fledged. Netscape had not developed into a real operating system competitor and might never have done so. As the D.C. Circuit explained, the relevant question is whether the targets "reasonably constitut[e] nascent threats."[86] Second (and implied by the first point), the target need not operate in the same market as the monopolist. By way of illustration, Netscape did not make operating systems, and therefore was not a participant in the relevant market of Intel-compatible PC operating systems.

Third, monopolizing conduct can take the form of collaboration rather than pure exclusion. The typical monopolization case focuses on exclusionary conduct that harms a rival, and here too *Microsoft* is exemplary. But Section 2 also reaches acquisitions and horizontal agreements, such as an agreement to divide markets.[87] A famous example is the consolidation of market power by Standard Oil.[88] If Microsoft had acquired Netscape rather than, say, acting to suppress browser distribution, that acquisition would violate Section 2. In fact, at one point, Microsoft apparently approached Netscape about buying or licensing Netscape's browser code[89] and later sought a market allocation arrangement in

[83] *Microsoft II*, 253 F.3d at 53–54 (assessing the perceived threat to Microsoft posed by Netscape's browser).

[84] Id. at 45.

[85] Id. at 46.

[86] Id. at 79.

[87] See, e.g., IV Phillip E. Areeda & Herbert Hovenkamp, Antitrust Law: An Analysis of Antitrust Principles and Their Application ¶ 912b, at 92 (4th ed. 2016) (concluding that the acquisition of a nascent rival "tends to maintain a monopoly by cutting off an avenue of future competition before it has a chance to develop. As a result, condemnation under § 2 is appropriate").

[88] Standard Oil Co. v. United States, 221 U.S. 1, 73–75 (1911) (agreeing with the court below that the 1899 consolidation of control in Standard Oil of New Jersey "operated to destroy the 'potentiality of competition' which otherwise would have existed"). The Supreme Court affirmed the lower court's conclusion that this conduct violated Sections 1 and 2 of the Sherman Act. Id. at 72–77.

[89] Plaintiffs' Joint Proposed Findings of Fact ¶ 64.1, United States v. Microsoft Corp. (*Microsoft I*), 84 F. Supp. 2d 9 (D.D.C. 1999) (No. 98-1221), https://www.justice.gov/atr/us-v-microsoft-proposed-findings-fact-1 (on file with the *Columbia Law Review*) (citing deposition testimony of a Microsoft executive describing the 1994 overture, prior to Microsoft's full recognition of the browser threat, to license Netscape browser software); id. ¶ 64.2 (quoting a Netscape executive's testimony that Microsoft had "offered a flat fee of a couple of

which Netscape would cease competing for PC-compatible browser business.[90]

As a further example, antitrust enforcers recently applied Section 2 to the acquisition of a nascent competitor. The drug maker Questcor made an unpatented blockbuster drug treatment for infantile spasms.[91] A European treatment posed a competitive threat to Questcor's monopoly, a threat that was merely nascent because the drug was not approved in the United States.[92] Questcor bought the U.S. rights to the European treatment, outbidding several other would-be acquirers.[93] The FTC and several states challenged, as unlawful monopoly maintenance, the defensive acquisition of a nascent competitor. Ultimately, Questcor agreed to pay $100 million and to license the acquired drug to another manufacturer to settle the case.[94]

Section 2 is a suitable vehicle for challenging consummated mergers. An antitrust enforcer need not block a transaction beforehand; waiting is fully within its power. Antitrust law has a statute of limitations, but it is directed to damages, not injunctive relief.[95] Laches—an unreasonable delay in bringing the suit—is generally understood to apply to private parties,[96] not the government.[97] Moreover, challenging a

million dollars to take us out of the game. And that would have killed our product in their space").

[90] Id. ¶ 67 (describing evidence of a June 1995 meeting in which Microsoft proposed that Netscape not develop a browser for Windows 95); see also *Microsoft I*, 84 F. Supp. 2d at 30–33 (describing efforts to "[d]issuade Netscape from [d]eveloping Navigator as a [p]latform"); Email from Bill Gates to Paul Maritz (May 31, 1995), https://www.justice.gov/sites/default/files/atr/legacy/2006/03/03/22.pdf [https://perma.cc/8VD9-MH3G] ("I think there is a very powerful deal of some kind we can do with Netscape. . . . I would really like to see something like this happen!!"). Microsoft also discussed internally the possibility of investing in Netscape. Id. ("Of course, over time, we will compete on the servers, but we can help them a lot in the meantime. We could even pay them money as part of the deal, buying some piece of them or something.").

[91] Complaint at 2, FTC v. Mallinckrodt ARD Inc., No. 1:17-cv-00120 (D.D.C. Jan. 18, 2017), https://www.ftc.gov/system/files/documents/cases/170118mallinckrodt_complaint_public.pdf [https://perma.cc/SG65-KUA5]. Questcor was subsequently acquired by Mallinckrodt. Id. at 4.

[92] Id. at 2–3.

[93] Id. at 9, 11–12.

[94] Mallinckrodt Ard Inc. (Questcor Pharmaceuticals), FTC, https://www.ftc.gov/enforcement/cases-proceedings/1310172/mallinckrodt-ard-inc-questcor-pharmaceuticals [https://perma.cc/6866-3LHK] (last updated July 14, 2017).

[95] 15 U.S.C. § 15b (2012) (establishing a four-year statute of limitations for suits seeking monetary damages); V Areeda & Hovenkamp, supra note 87, ¶ 1205b, at 309 ("[T]he four-year limitation applies only to damage suits, not to actions in equity.").

[96] See Int'l Tel. & Tel. Corp. v. Gen. Tel. & Elecs. Corp., 518 F.2d 913, 926 (9th Cir. 1975) ("We hold that . . . the defense of laches is available in [Clayton Act] § 16 suits . . . ."); V Areeda & Hovenkamp, supra note 87, § 1205b, at 309 ("The merged firm might have the defense of laches against a private suit . . . .").

[97] See *Int'l Tel. & Tel. Corp.*, 518 F.2d at 928 ("Laches cannot ordinarily be asserted against the sovereign."); United States v. Pennsalt Chems. Corp., 262 F. Supp. 101, 101

consummated transaction has a remedy—to undo the acquisition—that is closely connected to the nature of the unlawful conduct. This close connection sidesteps the concern, often raised in Section 2 cases, that a proposed remedy does not correspond closely enough to the established harm.[98]

The D.C. Circuit's language in *Microsoft* applies readily to today's online platforms: "[I]t would be inimical to the purpose of the Sherman Act to allow monopolists free rei[n] to squash nascent, albeit unproven, competitors at will—particularly in industries marked by rapid technological advance and frequent paradigm shifts."[99] As applied to the Instagram acquisition, the concern is that Facebook acquired a nascent, albeit unproven, competitor in social network services, thereby eliminating the risk of competition. Such an allegation, if ultimately supported by the facts, would be actionable under Section 2.

To make out a violation of Section 2, the enforcer would need to establish Facebook's monopoly power in a well-defined market[100]—for example, a market to provide social network services to users or a market for advertisements on social networks. The case would also entail an inquiry into anticompetitive effects[101]—for example, higher prices to advertisers and lower quality to users, in the form of more ads or less privacy protection. Facebook's demonstrated intent in acquiring Instagram, as established by documents or testimony, might furnish a further basis for inferring effect. The case might ultimately be strengthened and broadened if the facts showed that Facebook had engaged in a program of serial defensive acquisitions—for example, by purchasing the WhatsApp messaging service and perhaps other firms in order to neutralize the threat that they posed.[102]

(E.D. Pa. 1967) ("Laches is no defense in a suit by the government to vindicate a public right."); see also II Areeda & Hovenkamp, supra note 87, ¶ 320g, at 375 (describing this as the "usual proposition").

[98] See Massachusetts v. Microsoft Corp., 373 F.3d 1199, 1231 (D.C. Cir. 2004) (en banc) (describing the concern that a "drastic remedy, such as divestiture would be inappropriate if Microsoft's dominant position in the operating system market could not be attributed to its unlawful conduct").

[99] *Microsoft II*, 253 F.3d 34, 79 (D.C. Cir. 2001) (en banc) (per curiam).

[100] See United States v. Grinnell Corp., 384 U.S. 563, 570–71 (1966) (requiring, as an element of the monopolization offense, "the possession of monopoly power in [a] relevant market").

[101] *Microsoft II*, 253 F.3d at 58 ("[T]o be condemned as exclusionary, a monopolist's act must have an 'anticompetitive effect.' That is, it must harm the competitive *process* and thereby harm consumers.").

[102] This analysis is confined to Facebook's acquisitions. For an argument that Facebook has engaged in a pattern of false statements and deceptive conduct, thereby violating Section 2, see generally Dina Srinivasan, The Antitrust Case Against Facebook: A Monopolist's Journey Towards Pervasive Surveillance in Spite of Consumers' Preference for Privacy, 16 Berkeley Bus. L.J. 39 (2019).

In response, Facebook might be expected to argue—as CEO Mark Zuckerberg has in fact asserted—that Instagram would not have become a success were it not for Facebook's acquisition.[103] However, this argument ignores the fact that Instagram had strong alternative sources of support at its disposal, including venture capital funding and an acquisition offer, which Instagram spurned in favor of Facebook.[104] A second response is that Facebook, not competition, can best provide the security and privacy that consumers demand. However, accepting such an argument as a defense to an antitrust claim would defeat the fundamental policy animating the Sherman Act. As the Supreme Court has explained, the "Act reflects a legislative judgment that, ultimately, competition will produce not only lower prices but also better goods and services. . . . [T]he statutory policy precludes inquiry into the question whether competition is good or bad."[105]

Section 2 is not the only way to analyze the acquisition of a nascent competitor. An alternative framework is Section 7 of the Clayton Act, which prohibits mergers whose effect "may be substantially to lessen competition, or to tend to create a monopoly."[106] Section 7 is the customary legal tool for evaluating mergers. A particular acquisition might violate only Section 2, only Section 7, or both. The answer to one does not dictate the answer to the other.

The bread-and-butter Section 7 case is a merger of existing rivals, in which the merger is alleged to lessen competition in one or more well-defined markets in which both firms compete. Applied to Instagram, such an inquiry might focus on whether the merger increased concentration[107] or removed head-to-head competition as to photo-sharing services or social network services. A clear affirmative answer would support antitrust liability.

However, analysis of a nascent competitor within the traditional Section 7 framework tends to raise certain difficulties. There may be ambiguity about whether the acquirer and target compete (or competed) in the same market, a question that is important to establish a presumption of illegality in Section 7 challenges to horizontal mergers.[108] Relatedly, the traditional Section 7 framework tends to focus attention on current competition between existing rivals, whereas Section 2 focuses directly on the core competitive concern—removal of a nascent threat. Finally, although an anticompetitive effect need not be established with

[103] The Aspen Inst., A Conversation with Mark Zuckerberg, YouTube (June 26, 2019), https://www.youtube.com/watch?v=uHk2WfL5Gs4 (on file with the *Columbia Law Review*).

[104] See Swisher, supra note 69 (discussing acquisition offer from Twitter).

[105] Nat'l Soc'y of Prof'l Eng'rs v. United States, 435 U.S. 679, 695 (1978).

[106] 15 U.S.C. § 18 (2012).

[107] See United States v. Phila. Nat'l Bank, 374 U.S. 321, 363 (1963) (describing a presumption of illegality for horizontal mergers that significantly increase concentration).

[108] See id.

certainty in a Section 7 case,[109] judicial tolerance of uncertainty is expressed with greater clarity and prominence in the Section 2 context.[110] These differences tend to favor a Section 2 approach over a traditional Section 7 case.[111]

III. Disruptive Incumbents

A. *Reconciling Arrow and Schumpeter: Innovation in Adjacent Markets*

A second, and underappreciated, source of platform competition comes from disruptive incumbents. As discussed in Part I, Arrow and others have explained why we might expect competition to spur innovation. A second strand of the economic literature, traceable to the work of Schumpeter, points the other way by making a positive association between innovation and incumbency. This association has ex ante and ex post components.

Ex ante, the prospect of acquiring market power elicits innovation. For example, Google developed a new algorithmic approach to identifying important web content, displacing earlier search technologies.[112] Amazon built a superior online retail product over time.[113] These developments were motivated in part by an expectation of future profits. Schumpeter emphasized the role of such profits as a means to "lure capital on to untried trails" and thereby foster creative destruction.[114] These examples illustrate the Supreme Court's conclusion that the prospect of

[109] See, e.g., United States v. Baker Hughes Inc., 908 F.2d 981, 984 (D.C. Cir. 1990) ("Section 7 involves *probabilities*, not certainties or possibilities.").

[110] See supra note 86 and accompanying text.

[111] An alternative approach to Section 7 would do without the presumption of illegality for certain horizontal mergers, asking instead whether the transaction "tend[s] to create a monopoly." 15 U.S.C. § 18. This approach would roughly track the Section 2 inquiry in substance.

[112] See John Battelle, The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture 73–77, 151 (2005).

[113] See Letter from Jeffrey P. Bezos, Chief Exec. Officer, Amazon.com, Inc., to Shareholders (1997), https://ir.aboutamazon.com/static-files/589ab7fe-9362-4823-a8e5-901f6d3a0f00 [https://perma.cc/XT2R-PSSF] ("At this stage, we choose to prioritize growth because we believe that scale is central to achieving the potential of our business model.").

[114] Joseph A. Schumpeter, Capitalism, Socialism, and Democracy 89 (3d ed. 1950) ("[L]argest-scale plans could . . . not materialize at all if it were not known from the outset that competition will be discouraged by heavy capital requirements or lack of experience, or that means are available to discourage or checkmate it so as to gain the time and space for further developments."). Schumpeter argued that innovators require the means to "safeguard" investment through "insuring or hedging," id. at 88, and that monopoly is valuable as protection "against temporary disorganization of the market," id. at 103. This point is echoed in Peter Thiel, Zero to One: Notes on Startups, or How to Build the Future 33 (2014) ("[T]he promise of years or even decades of monopoly profits provides a powerful incentive to innovate.").

achieving a monopoly is valuable because "it induces risk taking that produces innovation and economic growth."[115]

The strong ex ante incentive to innovate identified by Schumpeter is consistent with the weak ex post innovation incentives identified by Arrow.[116] An entrant might work hard to attain a monopoly, and then, upon achieving it, turn to the quiet life. That consistency still leaves conflict at the level of antitrust or regulatory policies that facilitate the entry of outsiders. For example, prohibiting exclusionary conduct by a monopolist encourages entry by new competitors while simultaneously weakening the "pieces of armor" on which the incumbent's initial ex ante incentive might be partly based.[117] This tradeoff raises a question about how best to balance the two effects.[118]

Ex ante incentives are just one part of the Schumpeterian account. Schumpeter and others have argued that monopoly is also a potent platform for further innovation.[119] To be sure, some incumbents remain innovative even after they achieve a strong market position. For example,

[115] The Court stated:

> The mere possession of monopoly power, and the concomitant charging of monopoly prices, is not only not unlawful; it is an important element of the free-market system. The opportunity to charge monopoly prices—at least for a short period—is what attracts 'business acumen' in the first place; it induces risk taking that produces innovation and economic growth.

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

[116] See, e.g., Carl Shapiro, Competition and Innovation: Did Arrow Hit the Bull's Eye?, *in* The Rate and Direction of Inventive Activity Revisited 361, 401 (Josh Lerner & Scott Stern eds., 2012) (emphasizing consistency between Arrow's view that "a firm with a vested interest in the status quo has a smaller incentive than a new entrant to develop . . . new technology that disrupts the status quo" and Schumpeter's view that "the prospect of obtaining market power is a necessary reward to innovation").

[117] Schumpeter, supra note 114, at 89.

[118] For arguments that antitrust enforcement is desirable from an innovation standpoint because it improves the Arrovian incentive more than it suppresses the Schumpeterian incentive, see, for example, Hemphill & Wu, supra note 62, at 1211–12 n.137 (arguing that "[w]here self-entrenchment excludes an innovator . . . [the] negative effect on the innovative entrant" is likely to dominate); see also Christina Bohannan & Herbert Hovenkamp, Creation Without Restraint: Promoting Liberty and Rivalry in Innovation 245–50 (2011) (collecting examples in which dominant firms have slowed innovation by obstructing the market entry of innovative outsiders); Jonathan B. Baker, Beyond Schumpeter vs. Arrow: How Antitrust Fosters Innovation, 74 Antitrust L.J. 575, 583–88 (2007) (surveying empirical literature about the opposing effects).

[119] See Schumpeter, supra note 114, at 101 ("[T]here are advantages which, though not strictly unattainable on the competitive level of enterprise, are as a matter of effect secured only on the monopoly level . . . ."); F.M. Scherer, Schumpeter and Plausible Capitalism, 30 J. Econ. Literature 1416, 1418 (1992) ("Schumpeter went far beyond economists' long-accepted view that the *expectation* of a monopoly position . . . was necessary to make the venture worth while. Monopoly power *already held* also supported investment in technological progress."). In addition to Schumpeter, see, for example, Richard J. Gilbert & David M. G. Newbery, Preemptive Patenting and the Persistence of Monopoly, 72 Am. Econ. Rev. 514, 514 (1982) (offering a model of preemptive innovation by monopolists).

as discussed in Part I, leading platforms have aggressively adopted machine learning techniques to improve their matching algorithms and other aspects of their business. Incumbency may confer advantages in both the capacity and incentive to innovate. The advantages in capacity include superior personnel, greater financial resources, and the freedom to make long-term plans.[120] The incentive comes both from size—a large base over which to apply an improvement—and market power that allows the firm to appropriate the returns from further improvements.[121]

These ex post effects of incumbency run contrary to the Arrow replacement effect. However, there is an important and neglected point of reconciliation between the two perspectives. Arrow and Schumpeter coincide in their attitude toward innovative efforts *outside the home market* of the incumbent. Arrow's point is about innovation that cannibalizes the monopoly; the incentive to innovate in other markets is undiminished. Schumpeter's main focus, in its ex post component, is innovation (whether cannibalizing or not) that takes advantage of the incumbent's distinctive capacity. Thus, pursuing innovation outside the home market harnesses a variety of Schumpeterian advantages while avoiding the pitfalls of Arrow's replacement effect.[122]

This reconciliation is illustrated by leading platforms' aggressive forays outside of their home markets. For example, as noted in Part I, Amazon has built AWS into an important business selling storage and computing power to other firms.[123] Alphabet has undertaken an enormous effort to develop an autonomous vehicle. This research and commercialization effort is currently centered in Alphabet's Waymo subsidiary.[124] If successful, this project may upend the existing businesses of logistics, transportation services, and car manufacturing.[125]

[120] See Schumpeter, supra note 114, at 101–03 (discussing advantages of "brains," "higher financial standing," and "space . . . for long-range planning"). Franklin Fisher and Peter Temin usefully labeled this the "supply of innovations" argument (as distinct from the incentives-based "demand for innovations" argument). Franklin M. Fisher & Peter Temin, Returns to Scale in Research and Development: What Does the Schumpeterian Hypothesis Imply?, 81 J. Pol. Econ. 56, 57 (1973); see also Thiel, supra note 114, at 33 ("[M]onopolies can keep innovating because profits enable them to make the long-term plans and to finance the ambitious research projects that firms locked in competition can't dream of.").

[121] See, e.g., Baker, supra note 118, at 578 (explaining this incentive).

[122] Cf. Timothy F. Bresnahan & Shane Greenstein, Technological Competition and the Structure of the Computer Industry, 47 J. Indus. Econ. 1, 3 (1999) (describing divided technical leadership, wherein different firms furnish components of a platform, as an important source of innovative entry).

[123] See Amazon 2018 Annual Report, supra note 13, at 23–24 (reporting $7.3 billion in operating income and $25.7 billion in net sales from AWS in 2018).

[124] See Matthew DeBord, Waymo Could Be Worth as Much [as] $175 Billion—Here's a Brief History of the Google Car Project, Bus. Insider (Sept. 9, 2018), https://www.businessinsider.com/google-car-project-history-2018-8 [https://perma.cc/E98C-BKUU] (describing this effort).

[125] See id. (discussing Waymo's ambitions).

AWS and Waymo also illustrate a complementarity in production, whereby a large firm's core operations create capabilities that are profitably deployed elsewhere.[126] AWS began as an incidental byproduct of Amazon's effort in the early 2000s to improve certain aspects of its internal business processes.[127] The resulting improvements enabled Amazon to market its computing capabilities to other firms.[128] As for Waymo, machine learning is central to the development of its autonomous vehicle,[129] and Waymo has deployed deep-learning expertise developed in Google's core search business to solve certain technical challenges.[130] Moreover, Google serves as a supplier of machine learning software and specialized hardware to Waymo.[131]

Waymo further illustrates the freedom to make long-term plans emphasized by Schumpeter and others.[132] Waymo is a result of Alphabet's research lab, a modern-day version of the corporate research labs that played an important role in twentieth-century innovation.[133] The most

[126] See Paul Krugman, Robin Wells & Kathryn Graddy, Essentials of Economics 77 (2d ed. 2010) (defining complements in production). For example, natural gas is collected as a by-product of crude oil production, and sawdust (used in particleboard) is produced as a by-product of logging. The same term can refer instead to a variant of complementary demand in which multiple inputs are necessary to produce a product (such as fuel and airplanes for an airline). See, e.g., Daniel F. Spulber, Global Competitive Strategy 56 (2007) (using this definition).

A further supply-side complementarity is spare labor capacity, particularly scarce technical employees whose retention may require the leeway to spend part of their time on whatever interests them. In that case, the workers' preferences effectively lower the marginal cost of additional innovation. An example is the "20% time" historically granted by Google to pursue projects outside an employee's core responsibilities, leading to AdSense, Gmail, and Street View. See Christopher Mims, Google's "20% Time," Which Brought You Gmail and AdSense, Is Now as Good as Dead, Quartz (Aug. 16, 2013), https://qz.com/115831/googles-20-time-which-brought-you-gmail-and-adsense-is-now-as-good-as-dead [https://perma.cc/9LLK-VQSU].

[127] 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/QJ4U-LB8J] ("The internal teams at Amazon required a set of common infrastructure services everyone could access without reinventing the wheel every time, and that's precisely what Amazon set out to build—and that's when they began to realize they might have something bigger.").

[128] See id.

[129] See Dmitri Dolgov, Google I/O Recap: Turning Self-Driving Cars from Science Fiction into Reality with the Help of AI, Medium (May 8, 2018), https://medium.com/waymo/google-i-o-recap-turning-self-driving-cars-from-science-fiction-into-reality-with-the-help-of-ai-89dded40c63 [https://perma.cc/4D2M-8847] ("AI plays a crucial role in nearly every part of [Waymo's] self-driving system.").

[130] Id. (describing the "jump-start" Waymo gained by applying Google's deep-learning research to the problem of pedestrian detection).

[131] Id. ("At Waymo, we use the TensorFlow ecosystem and Google's data centers—including TPUs [a type of custom hardware]—to train our neural networks.").

[132] See supra note 120 and accompanying text.

[133] Early examples include General Electric (1901), DuPont (1902), Parke-Davis (1902), the Bell System (1911), and Kodak (1913). See George Basalla, The Evolution of

famous example is AT&T's Bell Labs, which simultaneously pursued basic science and innovations with a clear connection to AT&T operations, such as the development of the transistor.[134] It is not always clear—or, given the nature of the work, even knowable—whether such investments are rational from the standpoint of maximizing shareholder value. But their potential to alter the competitive conditions outside the incumbent's home market is undeniable.

B. *Platforms Targeting Platforms*

A special case of innovation and competition outside the home market is particularly relevant for our purposes: where one incumbent platform launches an attack on the core business of another. In recent years, there has been a remarkable variety of efforts by the leading tech firms to compete in one another's core businesses.[135] For example, Google has challenged Amazon in shopping starts (Google Shopping, among other efforts),[136] Facebook in social network services (Google+),[137] Apple in smartphone software (Android),[138] and Microsoft in productivity and operating system software (Google Docs, Gmail, Chrome).[139]

Technology 125–26 (1988) ("The first firms to engage in organized research were those whose technologies were closely linked to two areas of science that flourished in the late nineteenth century, chemistry and electricity.").

[134] See Tim Wu, The Master Switch: The Rise and Fall of Information Empires 104–07 (2010) (discussing, among other Bell Labs breakthroughs, the invention of the transistor). Bell researchers also discovered background radiation that helped to confirm the Big Bang theory. See James E. McClellan III & Harold Dorn, Science and Technology, *in* World History 374–75 (2d ed. 2006). Bell researcher Claude Shannon's work on information theory laid the groundwork for a revolution in digitization and signal processing. See Jon Gertner, The Idea Factory: Bell Labs and the Great Age of American Innovation 115–35 (2012) (discussing this work and its influence).

[135] See David S. Evans, Why the Dynamics of Competition for Online Platforms Leads to Sleepless Nights, but Not Sleepy Monopolies 22 (Aug. 23, 2017) (unpublished manuscript) (on file with the *Columbia Law Review*) ("Unlike the largest firms at previous points in time, these large Internet firms compete with each other across a range of products and services, despite each having gotten a toehold in the digital economy doing completely different things from one another.").

[136] See Claire Cain Miller & Stephanie Clifford, Google Struggles to Unseat Amazon as the Web's Most Popular Mall, N.Y. Times (Sept. 9, 2012), https://www.nytimes.com/2012/09/10/technology/google-shopping-competition-amazon-charging-retailers.html [https://perma.cc/KF4R-QFQB] ("Google is a search engine, not a store, but it is increasingly inching into e-commerce with products like its comparison-shopping service, Google Shopping.").

[137] See Farhad Manjoo, The Great Tech War of 2012, Fast Company (Oct. 19, 2011), https://www.fastcompany.com/1784824/great-tech-war-2012 [https://perma.cc/XV6P-D6GW].

[138] See id.

[139] Nick Wingfield, Windows 10 Signifies Microsoft's Shift in Strategy, N.Y. Times (July 19, 2015), https://www.nytimes.com/2015/07/20/technology/windows-10-signifies-microsofts-shift-in-strategy.html [https://perma.cc/A86L-RNRY] (attributing Microsoft's decision to offer a Windows 10 upgrade and a mobile version of Office for free to the fact

Such challenges are a potentially powerful source of platform competition and innovation. They harness the Schumpeterian capabilities identified in the previous section, without the incentives drag of the Arrow replacement effect. Indeed, as noted in the Introduction, an incumbent has an affirmative motivation to enter and compete in certain adjacent markets. In particular, an incumbent with market power has an affirmative preference for more intense competition in complementary products.[140] If the price of nails falls, demand for hammers increases. Thus, a producer of hammers has an incentive to arrange, if it can, a decrease in the price of nails.

One familiar application of the complementary demand effect is that manufacturers prefer more intense competition among distributors, and vice versa.[141] This preference helps to explain Google's investments to increase competition downstream in the provision of broadband internet service,[142] and Amazon's investments to increase competition upstream in book publishing.[143] But the point applies more broadly.

The complementary demand effect encourages entry on the margin that would be unprofitable for an ordinary firm. In the extreme case, an incumbent might be willing to introduce or sponsor competition that eliminates profits in the complementary market entirely. This incentive depends upon market power in the home market, in order to internalize the positive externality created by introducing new competition. Moreover, the larger the incumbent's share in the home market, the larger the effect.[144] Google sees a larger benefit from lower broadband prices than Bing. Amazon, compared to smaller retailers, sees a larger gain from intensified competition among publishers.[145]

that "[c]ompanies like Google have crept into Microsoft's business with free software and services subsidized by its huge advertising business").

[140] Goods are complements when a fall in one good's price increases demand for the other.

[141] See Cont'l Television, Inc. v. GTE Sylvania Inc., 433 U.S. 36, 56 & n.24 (1977) (discussing a manufacturer's interest in low retail prices to minimize the cost of distribution).

[142] Experimenting with New Ways to Make Broadband Better, Faster, and More Available, Google Pub. Pol'y Blog (Feb. 10, 2010), https://publicpolicy.googleblog.com/2010/02/experimenting-with-new-ways-to-make.html [https://perma.cc/NQ74-JA76] (describing Google's investment in new broadband infrastructure).

[143] See Kindle Direct Publishing, Amazon, https://kdp.amazon.com [https://perma.cc/2Q96-EMY3] (last visited Sept. 30, 2019). Moreover, in 2009, Amazon entered the traditional full-service publishing business with new imprints that hewed closely to the traditional model, including advances to authors, editorial services, and distribution to bookstores. See About Us, Amazon Publ'g, https://amazonpublishing.amazon.com/about-us.html [https://perma.cc/9FSW-VKF9] (last visited Sept. 30, 2019).

[144] Here, market share serves not as a proxy for market power but as a measure of the gains from introducing competition in a complementary business.

[145] Note that for increased competition upstream, the relevant question is the incumbent's share of purchases of the key input—in this case, books. A monopolist in a local output market, with only a small share of purchases of an input, might have little

One particularly successful instance of cross-market entry is Google's development of Android. In 2005, Google acquired Android Inc. and proceeded to make various investments to improve its smartphone operating system.[146] Following Apple's release of the iPhone in 2007,[147] Android development focused on touchscreen interfaces.[148] Android's release in the following year enabled handset manufacturers to offer a competitive alternative to the iPhone (and its iOS operating system) without developing their own software.[149] Reflecting the competition between Android and Apple, Google's CEO resigned from the Apple board.[150] Today, Android is the leading operating system by volume and offers the primary competition to the iPhone platform.[151]

The Android example illustrates several benefits of cross-market disruption. First, entry increases product variety and expands consumer choice. Apple and Android phones are differentiated in features and style.[152] Second, entry constrains prices by satisfying demand at a lower

incentive to introduce competition upstream. Thus, the incentives of two firms to introduce competition in one another's markets may be highly asymmetric.

[146] Farhad Manjoo, A Murky Road Ahead for Android, Despite Market Dominance, N.Y. Times (May 27, 2015), https://www.nytimes.com/2015/05/28/technology/personaltech/a-murky-road-ahead-for-android-despite-market-dominance.html [https://perma.cc/FA33-3F5D].

[147] See Jeremy W. Peters, Long-Awaited iPhone Goes on Sale, N.Y. Times (June 29, 2007), https://www.nytimes.com/2007/06/29/technology/29cnd-phone.html [https://perma.cc/C5D6-NN8N].

[148] Richard Gao, Android and Its First Purchasable Product, the T-Mobile G1, Celebrate Their 8th Birthdays Today, Android Police (Sept. 23, 2016), https://www.androidpolice.com/2016/09/23/android-first-purchasable-product-t-mobile-g1-celebrate-8th-birthdays-today [https://perma.cc/N7HM-BRKU] (noting touch-screen functionality of the first Android device); see also Michael Simon, 10 Years Ago We Met the World's First Android Phone, and It Didn't Have a Headphone Jack, PCWorld (Sept. 23, 2018), https://www.pcworld.com/article/3308157/first-android-phone-t-mobile-g1-10th-anniversary.html [https://perma.cc/6NZB-J2HJ] (describing the competition in 2018 between the iPhone and the first Android phone).

[149] 10 Years Later, Android Operating System Continues to Lead the Competition (Infographic), Dig. Info. World (Sept. 26, 2018), https://www.digitalinformationworld.com/2018/09/smartphone-operating-system-market-share.html [https://perma.cc/74H8-TKJS] ("Google's open approach to building a mobile platform . . . proved to be successful and it took less than three years for Android to become the number 1 platform in the booming smartphone market . . . with sales surpassing one billion for the first time in 2014.").

[150] John Quitner, Why Google's Schmidt Resigned from Apple's Board, Time (Aug. 3, 2009), https://content.time.com/time/business/article/0,8599,1914350,00.html [https://perma.cc/44F8-6UJ6].

[151] Mobile Operating System Market Share Worldwide, Statcounter Glob. Stats, https://gs.statcounter.com/os-market-share/mobile/worldwide [https://perma.cc/QF8P-5586] (last visited Sept. 30, 2019) (reporting 76% and 22% market shares for Android and Apple/iOS, respectively).

[152] See, e.g., Android 10, Android, https://www.android.com/android-10 [https://perma.cc/VS37-BWVA] (last visited Sept. 30, 2019) (highlighting Android advantages); Apple Switch Page, Apple, https://apple.com/switch [https://perma.cc/FJ4Z-BRNF] (last visited Sept. 30, 2019) (highlighting iPhone advantages).

price and placing downward pressure on iPhone prices.[153] Third, entry places pressure on firms in the targeted market to innovate in return. For example, after Android's success offering larger phones, Apple followed suit, abandoning its earlier resistance.[154]

Cross-market entry has also targeted the leading online platforms, seeking to provide competition for Google in search, Facebook in social network services, and Amazon in shopping. For example, Microsoft has made a heavy investment in search with Bing. Google attempted entry into social network services with Google+. And Google has taken on Amazon with a series of specialized product search and search advertising offerings.

The results have been mixed. First, consider Bing, which entered the search business in a big way in 2009. A massive investment, supported in part by a flow of user queries encouraged by Microsoft software products, made Bing a significant competitor in the United States, with a 25% share of queries as of April 2019.[155] Google's U.S. query share has remained steady at just over 60%;[156] globally, its share of queries is higher.[157]

Google's entry into social network services was regarded by Facebook as an existential threat and has been credited with causing Facebook to increase its focus on "reliability" and "user experience," as opposed to the "move fast and break things" approach that characterized its early history.[158] Ultimately, Google+ was a failure, despite Google's efforts to encourage its user base to adopt it.[159]

In its competition with Amazon (and others) to be the starting place for shopping starts, Google has deployed a variety of offerings. Starting in

[153] Android Pressures Apple on iPhone Pricing, Forbes (Apr. 28, 2011), https://www.forbes.com/sites/greatspeculations/2011/04/28/android-pressures-apple-on-iphone-pricing (on file with the *Columbia Law Review*).

[154] Jim Edwards, Steve Jobs Turned Out to Be Completely Wrong About the Key Reason People Like the iPhone, Bus. Insider (Sept. 12, 2014), https://www.businessinsider.com/steve-jobs-was-wrong-about-big-phones-2014-9 [https://perma.cc/QTJ5-QDAY] (noting Apple's change of heart, despite its earlier dismissal of larger phones).

[155] Share of Search Queries Handled by Leading U.S. Search Engine Providers as of April 2019, Statista, https://www.statista.com/statistics/267161/market-share-of-searchengines-in-the-united-states (on file with the *Columbia Law Review*) (last visited Sept. 30, 2019) (reporting comScore data).

[156] Id.

[157] See Daisuke Wakabayashi & Cecilia Kang, It's Google's Turn in Washington's Glare, N.Y. Times (Sept. 26, 2018), https://www.nytimes.com/2018/09/26/technology/google-conservatives-washington.html [https://perma.cc/CV9P-C4KY] (stating that Google "has 90 percent of the global search market").

[158] Antonio García Martínez, Chaos Monkeys: Obscene Fortune and Random Failure in Silicon Valley 285 (2016).

[159] Jon Brodkin, Google Doubles Plus Membership with Brute-Force Signup Process, Ars Technica (Jan. 22, 2012), https://arstechnica.com/gadgets/2012/01/google-doubles-plus-membership-with-brute-force-signup-process [https://perma.cc/2RUV-PY5L] (discussing a Google product redesign that made it difficult to create a Google account without signing up for Google+).

2007, Google introduced "universal" search, in which specialized search results were blended in among the other results on the search engine results page (SERP).[160] The product universals pointed to third-party merchants offering the product for sale. Google displayed its specialized shopping search results prominently. Prominence matters because the higher a link is displayed on the SERP, the more traffic it receives. For example, at some points, prominent placement was triggered by the presence of an Amazon product listing among the top organic results.[161] One effect was to encourage shopping starts from Google's search page.

Later, Google replaced the product universals with specialized product ads that, if clicked, take the user to a third-party merchant's website. Google's evolving efforts in shopping are a form of disruptive innovation that has provided important competition to Amazon. Notwithstanding these efforts, Amazon is the current leader in product searches.[162]

C. *Protecting "Punching Up"*

These efforts by leading platforms to compete in each other's core businesses are socially valuable. Their importance is heightened when the adjacent incumbent is a uniquely plausible competitor. Such unique status is more likely when the targeted business is occupied by a powerful, entrenched incumbent, and the necessary scale of entry is difficult to develop from scratch. The leading platforms—Google in search, Facebook in social network services, Amazon in shopping—all fit the bill.

The potential value of entry by an adjacent incumbent is worth protecting on the margin, particularly when the target's market power is highly durable. It is a reason for antitrust enforcement to tread carefully when it comes to platforms attacking platforms. Thus, an enforcer might decline to intervene, in the exercise of prosecutorial discretion, if a platform uses its strength in one business in order to more effectively compete in another against the targeted platform.

This suggestion, at first blush, might seem contrary to the usual intuition in antitrust enforcement that a powerful incumbent should be kept on a tight rein. However, it builds upon the recognition that certain conduct by a firm with market power, otherwise unlawful, is permissible in

[160] See Google: Universal Search, Search Engine Land, https://searchengineland.com/library/google/google-universal-search [https://perma.cc/5SPU-733H] (last visited Sept. 30, 2019).

[161] This point was disclosed in an FTC staff report that was inadvertently released in part. See Memorandum to the Federal Trade Commission, Bureau of Competition 130 n.136 (Aug. 8, 2012), https://www.benedelman.org/pdf/ftc-google-8aug2012.pdf (on file with the *Columbia Law Review*) [hereinafter Google Staff Memo].

[162] Jumpshot, Inc., The Competitive State of eCommerce Marketplaces: Data Report Q2 2018, at 17 (2018), https://go.jumpshot.com/rs/677-KZC-213/images/Jumpshot-Q2-Data-Report.pdf [https://perma.cc/H2NR-V8B4] (showing a 54/46 split favoring Amazon in 2018, compared to a 54/46 split favoring Google in 2015). Amazon has also become an increasingly important source of competition in product advertising. Id. at 20.

support of breaking into a new market. In particular, a firm with market power may tie a second product or service to one in which the firm enjoys market power, and defend on the ground that the conditional sale strengthens a bid for new entry.[163] Ultimately, a lighter touch may offer a realistic path forward when there are no or few alternatives.

One plausible candidate for a lighter touch is Google's conduct related to product universals. In its pursuit of shopping starts, Google has competed not only with Amazon but also with comparison shopping engines (CSEs). CSEs are specialized, domain-specific search engines that present and compare prices of a product offered on various websites. CSEs received significant traffic from appearing in Google's organic listings. Google's placement of product universals resulted in promotion above CSEs.[164] Google also demoted CSEs within its ordinary SERP results.[165]

The antitrust concern raised by this conduct was that Google allegedly preserved a dominant position in search and search advertising by impeding the growth of businesses that could develop into significant competitors. According to critics, promoting its own product universals at the expense of CSEs denied users adequate access to CSE results that were preferred by users and harmed the CSEs by starving them of exposure to users. In 2017, the European Commission fined Google several billion dollars for its product universal conduct,[166] a decision that is currently on appeal.

This alleged conduct was investigated by the FTC, but ultimately FTC staff recommended against challenging the conduct as an antitrust violation.[167] One important reason is that the conduct lacked a clear anticompetitive effect.[168] Prominent placement of product universals, like other efforts to add "answers" in addition to lists of websites, arguably improved the search results. CSEs are not merchants but intermediaries that lead to merchants. Google's inclusion of product universals directly on the SERP sped up the user's connection to a merchant. On this view,

[163] Jefferson Par. Hosp. Dist. No. 2 v. Hyde, 466 U.S. 2, 23 n.39 (1984) (citing United States v. Jerrold Elecs. Corp., 187 F. Supp. 545, 555–58 (E.D. Pa. 1960)) ("[T]ying may be permissible when necessary to enable a new business to break into the market.").

[164] Google Staff Memo, supra note 161, at 130 n.136 (summarizing testimony that "Google used the occurrence of [CSEs] at positions 1–3 in the web ranking to boost Google's product universal to position one"). The explanation given for this promotion was that a CSE's presence implied that a product universal would be relevant too, and that product universals were more useful than organic links to other CSEs (and hence presumably merited a higher position), an explanation that FTC staff concluded had "some force." Id. at 82.

[165] See id. at 28.

[166] See Google Shopping Decision, supra note 63, at 212.

[167] See Google Staff Memo, supra note 161, at 86 (recommending against challenging this conduct given "legal hurdles" and "strong procompetitive justifications," while regarding the question as "close").

[168] Id.

the product universals replaced CSEs that provided a low-quality user experience.[169] A further related impediment to enforcement is that U.S. antitrust law is reluctant to second-guess decisions about product design, for fear of false condemnations of beneficial product improvements.[170]

The foregoing analysis suggests a further justification for the FTC's nonenforcement against Google—that Google's conduct strengthened its ability to compete with Amazon for shopping starts. In other words, even if some consumers preferred CSEs and benefited from their prominent placement, this loss might be tolerable in order to promote the more important opportunity for Google to serve as a serious shopping competitor to Amazon.

Taking a step back from specific examples, as a general matter, today's competition among the leading tech firms is historically contingent and may be fragile. There is a risk that the leading firms might shift their strategy away from confrontation in favor of détente. As both a practical and legal matter, it is difficult to force firms to compete if they prefer to sit tight. Thus, there is reason to hesitate before disrupting the currently favorable equilibrium, lest we end up with less competition rather than more.

To be clear, inaction has a downside. It might result in collateral damage to firms, such as CSEs, caught in the crossfire between platforms. Nor may this possible harm be dismissed without inquiry as merely the result of ordinary competition, or as harm to competitors without any consequence to consumers. A further problem is that the competition enabled by a lighter touch might not succeed. The evidence about platforms attacking platforms, judged in terms of outcomes, paints a mixed picture. So we need to be clear-eyed about the likelihood that this will really work.

An important limiting principle, in considering whether to adopt a lighter touch, is that the firm under examination must be engaged in "punching up"—that is, attempting to compete with a strong platform in an adjacent market. Bing, Google+, and Google's product universals are good examples, as is Apple's recent insistence that apps utilizing a Google or Facebook login must also implement Apple's new login process.[171] By contrast, consider Google's placement of universals in the context of local search. The conduct at issue was roughly analogous to pro-

[169] For a contrary perspective, see Charles Duhigg, The Case Against Google, N.Y. Times (Feb. 20, 2018), https://www.nytimes.com/2018/02/20/magazine/the-case-against-google.html [https://perma.cc/Q3KJ-PMWV].

[170] *Microsoft II*, 253 F.3d 34, 65 (D.C. Cir. 2001) (en banc) (per curiam) ("As a general rule, courts are properly very skeptical about claims that competition has been harmed by a dominant firm's product design changes.").

[171] Updates to the App Store Review Guidelines, Apple (June 3, 2019), https://developer.apple.com/news/?id=06032019j [https://perma.cc/UP2M-NNA8].

duct universals.[172] Here, however, Google had no platform target on a par with Amazon. Where a firm under examination is punching down rather than up, the argument for a lighter touch presented here does not apply.

CONCLUSION

Nascent competitors and disruptive incumbents in adjacent markets are important sources of rivalry for the leading online platforms. Preserving these sources of competition has varying implications for antitrust policy and the role of antitrust enforcers. Nascent competitors require extra vigilance from enforcers to ensure that a far-seeing platform does not acquire the firm when its competitive significance is clear to the platform but not yet to enforcers. By contrast, competition from disruptive incumbents may be enhanced most effectively by adopting a measure of reserve.

[172] See Michel Luca, Tim Wu & Yelp Data Science Team, Is Google Degrading Search? Consumer Harm from Universal Search, Berkeley Law (July 2015), https://www.law.berkeley.edu/wp-content/uploads/2015/04/Luca-Wu-Yelp-Is-Google-Degrading-Search-2015.pdf [https://perma.cc/Y4EF-K28W] (arguing that Google gives inferior placement to superior, independent providers of specialized search).

WILL ARTIFICIAL INTELLIGENCE EAT THE LAW? THE RISE OF HYBRID SOCIAL-ORDERING SYSTEMS

*Tim Wu**

Software has partially or fully displaced many former human activities, such as catching speeders or flying airplanes, and proven itself able to surpass humans in certain contests, like Chess and Go. What are the prospects for the displacement of human courts as the centerpiece of legal decisionmaking? Based on the case study of hate speech control on major tech platforms, particularly on Twitter and Facebook, this Essay suggests displacement of human courts remains a distant prospect, but suggests that hybrid machine–human systems are the predictable future of legal adjudication, and that there lies some hope in that combination, if done well.

INTRODUCTION

Many of the developments that go under the banner of artificial intelligence that matter to the legal system are not so much new means of breaking the law but of bypassing it as a means of enforcing rules and resolving disputes.[1] Hence a major challenge that courts and the legal system will face over the next few decades is not only the normal challenge posed by hard cases but also the more existential challenge of supersession.[2]

Here are a few examples. The control of forbidden speech in major fora, if once the domain of law and courts, has been moving to algorithmic judgment in the first instance.[3] Speeding is widely detected and punished by software.[4] Much enforcement of the intellectual property

[*] Julius Silver Professor of Law & Technology, Columbia Law School. My thanks to Vince Blasi, Ryan Doerfler, Noah Feldman, Sarah Knuckey, David Pozen, Olivier Sylvain, and participants in the Columbia faculty workshop series.

[1] In this Essay, the broader meaning of "artificial intelligence" is used—namely a computer system that is "able to perform tasks normally requiring human intelligence" such as decisionmaking. Artificial Intelligence, Lexico, https://www.lexico.com/en/definition/artificial_intelligence [https://perma.cc/86XR-2JZ8] (last visited July 31, 2019).

[2] A small legal literature on these problems is emerging. See, e.g., Michael A. Livermore, Rule by Rules, *in* Computational Legal Studies: The Promise and Challenge of Data-Driven Legal Research (Ryan Whalen, ed.) (forthcoming 2019) (manuscript at 1–2), https://papers.ssrn.com/abstract=3387701 (on file with the *Columbia Law Review*); Richard M. Re & Alicia Solow-Niederman, Developing Artificially Intelligent Justice, 22 Stan. Tech. L. Rev. 242, 247–62 (2019); Eugene Volokh, Chief Justice Robots, 68 Duke L.J. 1135, 1147–48 (2019).

[3] See infra Part II.

[4] See Jeffrey A. Parness, Beyond Red Light Enforcement Against the Guilty but Innocent: Local Regulations of Secondary Culprits, 47 Willamette L. Rev. 259, 259 (2011) ("Automated traffic enforcement schemes, employing speed, and red light cameras, are increasingly used by local governments in the United States." (footnote omitted)).

laws is already automated through encryption, copy protection, and automated takedowns.[5] Public prices once set by agencies (like taxi prices) are often effectively supplanted by prices set by algorithm.[6] Blockchain agreements are beginning to offer an alternative mechanism to contract law for the forging of enforceable agreements.[7] Software already plays a role in bail determination and sentencing,[8] and some are asking whether software will replace lawyers for writing briefs, and perhaps even replace judges.[9]

Are human courts just hanging on for a few decades until the software gets better? Some might think so, yet at many points in Anglo American legal history, courts have been thought obsolete, only to maintain their central role. There are, it turns out, advantages to adjudication as a form of social ordering that are difficult to replicate by any known means.[10] This Essay predicts, even in areas in which software has begun to govern, that human courts[11] will persist or be necessarily reinvented. It predicts, however, that human–machine hybrids will be the first replacement for human-only legal systems, and suggests, if done right, that there lies real promise in that approach. The case study of content control on online platforms and Facebook's review board is used to support these descriptive and normative claims.

The prediction that courts won't wholly disappear may seem an easy one, but what's more interesting is to ask why, when software is "eating" so many other areas of human endeavor. Compared with the legal system, software has enormous advantages of scale and efficacy of enforcement. It might tirelessly handle billions if not trillions of decisions in the time it takes a human court to decide a single case. And even more importantly, the design of software can function as an ex ante means of ordering that does not suffer the imperfections of law enforcement.[12]

But human courts have their own advantages. One set of advantages, more obvious if perhaps more fragile, is related to procedural fairness. As between a decision made via software and court adjudication, the latter, even if delivering the same results, may yield deeper acceptance and

[5] See infra Part I.

[6] See Jessica Leber, The Secrets of Uber's Mysterious Surge Pricing Algorithm, Revealed, Fast Company (Oct. 29, 2015), https://www.fastcompany.com/3052703/the-secrets-of-ubers-mysterious-surge-pricing-algorithm-revealed [https://perma.cc/H7MB-SC8T].

[7] See Eric Talley & Tim Wu, What Is Blockchain Good for? 5–9 (Feb. 28, 2018) (unpublished manuscript) (on file with the *Columbia Law Review*).

[8] See Algorithms in the Criminal Justice System, Elec. Privacy Info. Ctr., https://epic.org/algorithmic-transparency/crim-justice/ [https://perma.cc/KY2D-NQ2J] (last visited Apr. 23, 2019).

[9] See Volokh, supra note 2, at 1144–48, 1156–61.

[10] See Lon L. Fuller, The Forms and Limits of Adjudication, 92 Harv. L. Rev. 353, 357 (1978).

[11] This Essay uses "courts" to refer to any adjudicative body, public or private, that resolves a dispute after hearing reasoned argument and explains the basis of its decision.

[12] Lawrence Lessig, Code: Version 2.0, at 124–25 (2006).

greater public satisfaction.[13] In the future, the very fact of human decision—especially when the stakes are high—may become a mark of fairness.[14] That said, society has gotten used to software's replacement of humans in other areas, such as the booking of travel or the buying and selling of stocks, so this advantage may be fragile.

A second, arguably more lasting advantage lies in human adjudication itself and its facility for "hard cases" that arise even in rule-based systems. Most systems of social ordering consist of rules, and a decisional system that was merely about obeying rules might be replaced by software quite easily. But real systems of human ordering, even those based on rules, aren't like that.[15] Instead, disputes tend to be comprised of both "easy cases"—those covered by settled rules—and the aforementioned "hard cases"—disputes in which the boundaries of the rules become unclear, or where the rules contradict each other, or where enforcement of the rules implicates other principles.[16] There is often a subtle difference between the written rules and "real rules," as Karl N. Llewellyn put it.[17] Hence, a software system that is instructed to "follow the rules" will produce dangerous or absurd results.

Justice Cardozo argued that the judicial process "in its highest reaches is not discovery, but creation . . . [by] forces . . . seldom fully in consciousness."[18] Better results in hard cases may for a long time still depend instead on accessing something that remains, for now, human—that something variously known as moral reasoning, a sensitivity to evolving norms, or a pragmatic assessment of what works. It is, in any case, best expressed by the idea of exercising "judgment." And if the courts do indeed have a special sauce, that is it.

It is possible that even this special sauce will, in time, be replicated by software, yielding different questions.[19] But as it stands, artificial intelligence (AI) systems have mainly succeeded in replicating human decisionmaking that involves following rules or pattern matching—chess

[13] See generally E. Allan Lind & Tom R. Tyler, The Social Psychology of Procedural Justice (Melvin J. Lerner ed., 1988) (providing a classic overview of the advantages human courts have over software because of procedural fairness considerations in human courts, potentially ignored by software, that lead to acceptance of their decisions by the public).

[14] See Aziz Z. Huq, A Right to a Human Decision, 105 Va. L. Rev. (forthcoming 2020) (manuscript at 21–22), https://ssrn.com/abstract=3382521 (on file with the *Columbia Law Review*) (explaining that concerns about transparency have led some to demand human decisions).

[15] This refers to a caricatured version of H.L.A. Hart's view of what law is. See H.L.A. Hart, The Concept of Law 2–6 (Peter Cane, Tony Honoré & Jane Stapleton eds., 2d ed. 1961).

[16] Cf. Ronald Dworkin, Hard Cases, *in* Taking Rights Seriously 81, 81 (1978) (stating that when "hard cases" fall on the edge of clear rules, judges have the discretion to decide the case either way).

[17] Karl N. Llewellyn, The Theory of Rules 72–74 (Frederick Schauer ed., 2011).

[18] Benjamin N. Cardozo, The Nature of the Judicial Process 166–67 (1921).

[19] See Huq, supra note 14, at 18; Volokh, supra note 2, at 1166–67, 1183–84.

and Jeopardy! are two examples.[20] It would risk embarrassment to argue that machines will *never* be able to make or explain reasoned decisions in a legal context, but the challenges faced are not trivial or easily overcome.[21] And even if software gets better at understanding the nuances of language, it may still face the deeper, jurisprudential challenges described here. That suggests that, for the foreseeable future, software systems that aim to replace systems of social ordering will succeed best as human–machine hybrids, mixing scale and efficacy with human adjudication for hard cases. They will be, in an older argot, "cyborg" systems of social ordering.[22]

When we look around, it turns out that such hybrid systems are already common. Machines make the routine decisions while leaving the hard cases for humans. A good example is the flying of an airplane, which, measured by time at the controls, is now mostly done by computers, but sensitive, difficult, and emergency situations are left to a human pilot.[23]

The descriptive thesis of this Essay is supported by a case study of content control (the control of hate speech, obscenity, and other speech) on online platforms like Facebook and Twitter. Despite increasing automation, the generation of hard questions has yielded the development, by the major platforms, of deliberative bodies and systems of appeal, such as Facebook's prototype content review board, designed to rule on the hardest of speech-related questions. In the control of online speech, and in the autopilot, we may be glimpsing, for better or worse, the future of social ordering in advanced societies.

While automated justice may not sound appealing on its face, there is some real promise in the machine–human hybrid systems of social ordering described here. At their best, they would combine the scale and

[20] See John Markoff, Computer Wins on 'Jeopardy!': Trivial, It's Not, N.Y. Times (Feb. 16, 2011), https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html (on file with the *Columbia Law Review*).

[21] See Drew McDermott, Why Ethics Is a High Hurdle for AI 2 (2008), http://www.cs.yale.edu/homes/dvm/papers/ethical-machine.pdf [https://perma.cc/PZ7Z-5QFN]; Adam Elkus, How to Be Good: Why You Can't Teach Human Values to Artificial Intelligence, Slate (Apr. 20, 2016), https://slate.com/technology/2016/04/why-you-cant-teach-human-values-to-artificial-intelligence.html [https://perma.cc/4LR9-Q2GH]. But see IBM's project debater, a program that presents arguments on one side of an issue, and thereby might be thought to replicate aspects of lawyering. Project Debater, IBM, https://www.research.ibm.com/artificial-intelligence/project-debater/ [https://perma.cc/4A2W-XSB9] (last visited July 31, 2019).

[22] The Merriam-Webster dictionary defines "cyborg" as "a bionic human," meaning a being comprised of mixed human and machine elements, like the fictional character "Darth Vader" from the twentieth-century film series *Star Wars*. See Cyborg, Merriam-Webster Dictionary, https://www.merriam-webster.com/dictionary/cyborg [https://perma.cc/FD6A-UWZ2] (last visited Aug. 31, 2019).

[23] Reem Nasr, Autopilot: What the System Can and Can't Do, CNBC (Mar. 27, 2015), https://www.cnbc.com/2015/03/26/autopilot-what-the-system-can-and-cant-do.html [https://perma.cc/Q7CT-GBLN].

effectiveness of software with the capacity of human courts to detect errors and humanize the operation of the legal system. As Lon Fuller put it, human courts are "a device which gives formal and institutional expression to the influence of reasoned argument in human affairs."[24] What might the court system look like without scaling problems—if routine decisions went to machines, reducing the court's own workload as a major factor in decisionmaking? To be sure, what could arise are inhumane, excessively rule-driven systems that include humans as mere tokens of legitimization,[25] but hopefully we can do better than that.

This Essay provides advice both for designers of important AI-decision systems and for government courts. As for the former, many AI systems outside of the law have aspired to a complete replacement of the underlying humans (the self-driving car,[26] the chess-playing AI[27]). But when it comes to systems that replace the law, designers should be thinking harder about how best to combine the strengths of humans and machines, by understanding the human advantages of providing a sense of procedural fairness, explainability, and the deciding of hard cases. That suggests the deliberate preservation of mechanisms for resort to human adjudication (either public or private) as part of a long-term, sustainable system.

Human courts, meanwhile, should embark on a greater effort to automate the handling of routine cases and routine procedural matters, like the filing of motions. The use of intelligent software for matters like sentencing and bail—decisions with enormous impact on people's lives—seems exactly backward. The automation of routine procedure might help produce both a much faster legal system and also free up the scarce resource of highly trained human judgment to adjudicate the hard cases, or to determine which are the hard cases. Anyone who has worked in the courts knows that the judiciary's mental resources are squandered on thousands of routine matters; there is promise in a system that leaves judges to do what they do best: exercising judgment in the individual case, and humanizing and improving the written rules. This also implies that judges should seek to cultivate their comparative advantage, the exercise of human judgment, instead of trying to mimic machines that follow rules.[28]

[24] Fuller, supra note 10, at 366.

[25] This is a concern expressed in Re & Solow-Niederman, supra note 2, at 246–47.

[26] Alex Davies, The WIRED Guide to Self-Driving Cars, WIRED (Dec. 13, 2018), https://www.wired.com/story/guide-self-driving-cars/ [https://perma.cc/7P3A-5NLE].

[27] Natasha Regan & Matthew Sadler, DeepMinds's Superhuman AI Is Rewriting How We Play Chess, WIRED (Feb. 3, 2019), https://www.wired.co.uk/article/deepmind-ai-chess [https://perma.cc/WF5Z-VEV7].

[28] Cf. Kathryn Judge, Judges and Judgment: In Praise of Instigators, 86 U. Chi. L. Rev. (forthcoming 2019) (manuscript at 1–2), https://ssrn.com/abstract=3333218 (on file with the *Columbia Law Review*) (describing Judge Richard Posner's rejection of such machine-like jurisprudence).

Part I frames software decisionmaking among its competitors as a system of social ordering. Part II introduces the case study of Twitter and Facebook's handling of hate speech, focusing on the evolution of online norms, and the subsequent adoption of a hybrid human–software system to control speech. Part III assesses, from a normative perspective, the hybrid systems described in Part II.

I. BACKGROUND

In his book *Empire and Communications*, Harold Innes sought to characterize civilizations by their primary medium of communication.[29] The oral tradition of the ancient Greeks, he argued, informed the development of Greek philosophy; the Egyptian civilization changed as it transitioned from stone engraving to papyrus; medieval thought was shaped by the codex, and so on.[30] For our purposes, we need a different taxonomy of civilizations, one that characterizes societies not by medium but by how they make important public decisions (or, in Fuller's phrase, accomplish "social ordering").[31] By this I mean decisions that are both important and of public consequence, that define citizens' relationships with each other.

Under this way of seeing things, civilizations and societies really do differ. One axis is the divide between private and public bodies. Another is how much is governed by social norms as opposed to positive law.[32] Ordering might be more or less centralized; and there is the method of decision itself, which, as Fuller suggested, might be adjudicative, legislative, or accomplished by contractual negotiation.[33] I will not bother to pretend that the lines I have mentioned are the only ways you might imagine the division.[34]

This broader view demonstrates that decisional systems are in an implicit competition. Matters may drift between private and public decisionmaking, or between norms and law, and decisions can become more centralized or decentralized. Over the last 200 years, in the United States and commonwealth countries, a decentralized common law has been

[29] See generally Harold Innes, Empire and Communications (David Godfrey ed., 1950) (arguing that communication provides crucial insight into a civilization's organization and administration of its government, and comparing various civilizations including Egypt and the Roman Empire based on their communication).

[30] See id. at 5.

[31] See Fuller, supra note 10, at 357.

[32] See Robert C. Ellickson, Order Without Law: How Neighbors Settle Disputes 1–11 (1994) (discussing how order is often achieved without law).

[33] See Fuller, supra note 10, at 363 (arguing that "adjudication should be viewed as a form of social ordering, as a way in which the relations of men to one another are governed and regulated").

[34] There is, for example, also the question of how centralized or decentralized the systems of order are. Lawrence Lessig divided the universe of regulative forms into four: law, code, norms, and markets. Lessig, supra note 12, at 124–25.

somewhat (though not fully) displaced by a more centralized statutory law, and then further displaced by regulations and the administrative state.[35] Matters once thought purely private, like the firing of an employee or one's conduct in the workplace, have become subjects of public decision, while others once public, like the control of obscenity and other forbidden speech, are now mainly the province of private institutions. There is much complementarity in social ordering: a murderer may be shamed, fired, and imprisoned. But there is also competition, as for example, when new laws "crowd out" longstanding norms.

That idea that systems of public ordering might compete (or complement each other) is not new,[36] but what *is* new is the arrival of software and artificial intelligence as a major modality of public decisionmaking. As first predicted by Lawrence Lessig, what might have been thought to be important public decisions have either been displaced or are beginning to be displaced by software, in whole or in part.[37] It is a subtle displacement, because it is both private and unofficial, and advancing slowly, but it is happening nonetheless.

That idea of being ruled by intelligent software may sound radical but, as suggested in the Introduction, it is not hard to find examples in which software accomplishes what might previously be described as public decisionmaking. A good example is the dissemination and reproduction of expressive works. The posting of copyrighted works on YouTube and other online video sites was once directly and actively governed by section 512 of the copyright code.[38] In a technical sense the law still governs, but over the last decade sites like YouTube have begun using software (named "Content ID") to intelligently and proactively take down copyrighted works.[39] This understanding, implemented in code, was undertaken in the shadow of the law, but it is not compelled by it, and the decisions made by the software are now more important than the law. In the criminal law, software has become an aid to decisionmaking, and sometimes the decisionmaker in some jurisdictions, for matters like

[35] See Lawrence Friedman, A History of American Law 253–78, 503–15 (3d ed. 2005).

[36] See, e.g., Emanuela Carbonara, Law and Social Norms, *in* 1 The Oxford Handbook of Law & Economics 466, 475–80 (noting that while legal norms can reinforce social norms by "bending them towards the law when discrepancy exists and favoring their creation where social norms do not exist," legal regulation can also "destroy existing social norms"); Yoshinobu Zasu, Sanctions by Social Norms and the Law: Substitutes or Complements?, 36 J. Legal Stud. 379, 379–82 (2007) (discussing whether informal sanctions imposed through social norms are in competition with, or complement, the formal sanctions of the law).

[37] Lessig, supra note 12, at 125–37.

[38] 17 U.S.C. § 512 (2012).

[39] How Content ID Works, YouTube Help, https://support.google.com/youtube/answer/2797370?hl=en [https://perma.cc/CLD9-L2VK] [hereinafter YouTube Help, How Content ID Works] (last visited July 31, 2019).

setting bail or sentencing.[40] And as we shall see in much more detail below, the control of forbidden speech in major online fora is heavily dependent on software decisions.

To be sure, software remains in the early stages of replacing the law, and much remains completely outside software's ambit. But let us assume that software is at least beginning to be the method by which at least some decisions once made by the law are now made.[41] If that is true, then the best glimpse of what the future will hold lies in the systems that control offensive, hateful, and harmful speech online.

II. THE CASE STUDY: FACEBOOK, TWITTER, AND HEALTHY SPEECH ENVIRONMENTS

This Part provides a case study of the migration of Facebook and Twitter toward a norm of healthy speech environments and their implementation of such norms in hybrid systems of code and human judgment.

A. *The Evolution of Online Speech Norms from the 1990s Through 2016*

When the first online communities began emerging in the late 1980s and early 1990s, a widespread and early aspiration was the creation of public platforms that were "open and free" in matters of speech.[42] That desire reflected, in part, the "cyber-libertarian" tendencies among the pioneers of online technologies of that era.[43] The World Wide Web, which became popular in the early 1990s, was the chief enabling technology for the promise of a non-intermediated publishing platform for the masses. To a degree rarely, if ever, attempted in human history, the web and its major fora and platforms adhered to an "open and free" philosophy.[44] The Usenet, an early and public online discussion forum,

[40] Algorithms in the Criminal Justice System, supra note 8 (listing different states' uses of algorithmic tools for sentencing, probation, and parole decisions).

[41] The key to accepting this conclusion is to accede to the premise that the software is making decisions, which some may dispute. Some might ask if the software is really "deciding," as opposed to the programmer of the algorithm. I address these complications in Tim Wu, Machine Speech, 161 U. Pa. L. Rev. 1495 (2013).

[42] Tim Wu, The Attention Merchants: The Epic Scramble to Get Inside Our Heads 252 (2016) [hereinafter Wu, Attention Merchants].

[43] See Tim Wu & Jack Goldsmith, Who Controls the Internet: Illusions of a Borderless World 1–10 (2006).

[44] Even in the 1990s, the online communities that experimented with purely laissez faire speech platforms ran into problems linked to trolling and abuse, and very few of the communities were completely without rules. See Wu, Attention Merchants, supra note 42, at 276–88. It is also true that, by the early 2000s, the law and courts began to demand compliance with their laws, including the copyright laws, drug laws, laws banning child pornography, and so on. See Wu & Goldsmith, supra note 43, at 65–85.

allowed any user to create their own forum on any topic.[45] The famous online chatrooms of the 1990s were largely uncensored.[46] MySpace, the most popular social networking platform before Facebook, allowed its users to use any name they wanted, and to say almost anything they wanted.[47]

The "open and free" ideal was aided by the enactment of Section 230 of the Communications Decency Act of 1996.[48] The law, which granted platform owners immunity from tort for content posted on their platforms, was described as a "good Samaritan" law to protect sites trying to take down offensive content.[49] In practice, and in judicial interpretation, section 230 granted blanket immunity to all online platforms, both good Samaritans and bad, thereby protecting those who followed an "anything goes" mentality.[50]

The "open and free" speech ideal remained an aspired-to norm for the first twenty years of the popular internet. But under pressure, it began to change decisively over the years 2016 and 2017.[51] It has been replaced with a widespread if not universal emphasis among the major platforms—especially Twitter and Facebook—on creating "healthy" and "safe" speech environments online.[52] To be sure, the change in norms

[45] See Sandra L. Emerson, Usenet: A Bulletin Board for Unix Users, Byte, Oct. 1983, at 219, 219, https://archive.org/stream/byte-magazine-1983-10/1983_10_BYTE_08-10_UNIX#page/n219/mode/2up (on file with the *Columbia Law Review*).

[46] See Wu, Attention Merchants, supra note 42, at 202–05; see also EJ Dickson, My First Time with Cybersex, Kernel (Oct. 5, 2014), https://kernelmag.dailydot.com/issue-sections/headline-story/10466/aol-instant-messenger-cybersex/ [https://perma.cc/98EC-6F3H] (recounting experiences with cybersex as a ten-year-old).

[47] See Saul Hansell, For MySpace, Making Friends Was Easy. Big Profit Is Tougher., N.Y. Times (Apr. 23, 2006), https://www.nytimes.com/2006/04/23/business/yourmoney/23myspace.html (on file with the *Columbia Law Review*) (describing MySpace as "very open to frank discussion, provocative images and links to all sorts of activities" including profiles maintained by *Playboy* magazine and porn star Jenna Jameson); Michael Arrington, MySpace Quietly Begins Encouraging Users to Use Their Real Names, TechCrunch (Dec. 17, 2008), https://techcrunch.com/2008/12/17/myspace-quietly-begins-encouraging-users-to-use-their-real-names/ [https://perma.cc/CKR6-MLYZ] (noting pre-2009 anonymity of MySpace profiles).

[48] Communications Decency Act of 1996, 47 U.S.C. § 230 (2012) (stating, among other things, that "[i]t is the policy of the United States . . . to preserve the vibrant and competitive free market that presently exists for the Internet and other interactive computer services").

[49] See Andrew M. Sevanian, Section 230 of the Communications Decency Act: A "Good Samaritan" Law Without the Requirement of Acting as a "Good Samaritan," 21 UCLA Ent. L. Rev. 121, 125 (2014).

[50] See Danielle Keats Citron & Benjamin Wittes, The Internet Will Not Break: Denying Bad Samaritans Section 230 Immunity, 86 Fordham L. Rev. 401, 413 (2017) ("An overbroad reading of the [Communications Decency Act] has given online platforms a free pass to ignore illegal activities, to deliberately repost illegal material, and to solicit unlawful activities while ensuring that abusers cannot be identified.").

[51] See infra text accompanying notes 61–63.

[52] See infra notes 61–63 and accompanying text.

has never been explicitly stated as such; but it is hard to deny the change in emphasis is not also a change in substance. We might put it this way: If the major American online platforms once (from the 1990s through the mid-2010s) tended to follow speech norms that generally resembled the First Amendment's, the new focus on healthy speech and acceptance of the concept of harmful speech is far closer to the European speech tradition and its bans on hate speech.[53]

What changed? The mid-2010s shift in online speech norms on major platforms can be understood as reflecting three major developments. The first has been the relative success of a broader social movement stressing the importance of "safe" environments, reflected most strongly at American college campuses in the 2010s.[54] Those norms began to spill over into increasingly strong critiques of the major internet speech platforms. By the mid-2010s, journalists and civil rights groups, for example, heavily criticized Twitter and Facebook for tolerating attacks on women and historically disadvantaged groups and thereby creating "toxic" spaces for its users.[55] As a Buzzfeed journalist wrote in 2016, "Twitter is as infamous today for being as toxic as it is famous for being revolutionary."[56]

A second reason was a political concern: a widespread perception that the platforms had tolerated so much dissemination of hateful speech, foreign interference with elections, atrocity propaganda, and hoaxes as to become a threat to democratic institutions. This critique

[53] See Jeremy Waldron, The Harm in Hate Speech 6–17 (2012) (summarizing legal and philosophical differences between European and American speech traditions).

[54] In 2015, a large survey found about 71% of college entrants agreed that "colleges should prohibit racist/sexist speech on campus." Kevin Eagan, Ellen Bara Stolzenberg, Abigail K. Bates, Melissa C. Aragon, Maria Ramirez Suchard & Cecilia Rios-Aguilar, Higher Educ. Research Inst., The American Freshman: National Norms Fall 2015, at 47 (2016), https://www.heri.ucla.edu/monographs/TheAmericanFreshman2015.pdf [https://perma.cc/82CJ-T53B].

[55] See, e.g., Emily Dreyfuss, Twitter Is Indeed Toxic for Women, Amnesty Report Says, WIRED (Dec. 18, 2018), https://www.wired.com/story/amnesty-report-twitter-abuse-women/ [https://perma.cc/2NXS-SULS] (detailing high rates of abusive tweets directed toward women journalists and politicians); Robinson Meyer, The Existential Crisis of Public Life Online, Atlantic (Oct. 30, 2014), https://www.theatlantic.com/technology/archive/2014/10/the-existential-crisis-of-public-life-online/382017/ [https://perma.cc/7845-2PHH] (criticizing Twitter's lack of response to Gamergate); Hamza Shaban & Taylor Telford, Facebook and Twitter Get an Avalanche of Criticism About Russian Interference, L.A. Times (Dec. 18, 2018), https://www.latimes.com/business/technology/la-fi-tn-facebook-twitter-20181218-story.html [https://perma.cc/KS68-44S9] (describing the NAACP's criticism of Facebook for "the spread of misinformation and the utilization of Facebook for propaganda promoting disingenuous portrayals of the African American community").

[56] Charlie Warzel, "A Honeypot for Assholes": Inside Twitter's 10-Year Failure to Stop Harassment, BuzzFeed News (Aug. 11, 2016), https://www.buzzfeednews.com/article/charliewarzel/a-honeypot-for-assholes-inside-twitters-10-year-failure-to-s [https://perma.cc/V92P-NB7N].

emerged strongly after the 2016 election.[57] Relatedly, outside the United States over this period, Facebook faced heated accusations that its site was used to organize and promote violence in countries like Myanmar, Sri Lanka, and India.[58]

A final development was the ability, given consolidation in the speech platform market, for a limited number of platforms—Twitter, Facebook, Google—to have system-wide effects. (These platforms, all private actors, are of course unconstrained by constitutional norms.[59]) To be sure, there are some platforms, like 4chan, that remain devoted to the older laissez faire norm,[60] and specialized sites, like pornographic sites, that obviously take different views of sex and nudity. But by 2016, the major platforms, surely comprising most of the online speech in the world, had all effectively moved to treat hateful speech as potentially "violent," an "attack," and subject to removal.[61] The new norms of online speech are codified in the lengthy and highly specific content rules kept by Facebook, Google, and YouTube.[62] Simply put, they now regard many categories of speech as subject to removal, from the more easily defined (videos of suicide attempts, child pornography) to the more ambiguous

[57] See, e.g., Alexis C. Madrigal, What Facebook Did to American Democracy, Atlantic (Oct. 12, 2017), https://www.theatlantic.com/technology/archive/2017/10/what-facebook-did/542502/ [https://perma.cc/8AV5-AE3B] (chronicling Facebook's role in the 2016 elections and concluding that the "roots of the electoral system—the news people see, the events they think happened, the information they digest—had been destabilized").

[58] See, e.g., Vindu Goel & Shaikh Azizur Rahman, When Rohingya Refugees Fled to India, Hate on Facebook Followed, N.Y. Times (June 14, 2019), https://www.nytimes.com/2019/06/14/technology/facebook-hate-speech-rohingya-india.html (on file with the *Columbia Law Review*); Paul Mozur, A Genocide Incited on Facebook, with Posts from Myanmar's Military, N.Y. Times (Oct. 15, 2018), https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html (on file with the *Columbia Law Review*); Amalini De Sayrah, Opinion, Facebook Helped Foment Anti-Muslim Violence in Sri Lanka. What Now?, Guardian (May 5, 2018), https://www.theguardian.com/commentisfree/2018/may/05/facebook-anti-muslim-violence-sri-lanka [https://perma.cc/Y4X2-YCAG].

[59] See, e.g., Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1930 (2019) ("[W]hen a private entity provides a forum for speech, the private entity is not ordinarily constrained by the First Amendment because the private entity is not a state actor. The private entity may thus exercise editorial discretion over the speech and speakers in the forum.").

[60] See generally Rules, 4chan, https://www.4chan.org/rules [https://perma.cc/MSQ7-PN7N] (last visited July 30, 2019) (designating spaces where racism, pornography, and grotesque violence are allowed).

[61] See, e.g., Community Standards, Facebook, https://www.facebook.com/communitystandards/ [https://perma.cc/D27N-XJEY] (last visited July 30, 2019); Hate Speech Policy, YouTube Help, https://support.google.com/youtube/answer/2801939?hl=en [https://perma.cc/AZD2-VH4V] (last visited July 30, 2019).

[62] See supra note 61.

(hate speech, dehumanizing speech, advocacy of violence or terrorism).[63]

The easiest way to see the change in norms is by observing the changes in language used by representatives of the major companies. In 2012, Twitter executives had described the firm as belonging to "the free speech wing of the free speech party" and suggested that, in general "we remain neutral as to the content."[64] Alexander Macgillivray, Twitter's general counsel at the time, regularly litigated subpoena requests, telling the press that "[w]e value the reputation we have for defending and respecting the user's voice . . . . We think it's important to our company and the way users think about whether to use Twitter, as compared to other services."[65]

In contrast, by the later 2010s, Twitter had begun to emphasize "health" and "safety" as primary concerns.[66] In an interview, Twitter CEO Jack Dorsey suggested the "free speech wing" quote "was never a mission of the company" and that "[i]t was a joke, because of how people found themselves in the spectrum."[67] And, as the official Twitter blog stated in 2017:

> Making Twitter a safer place is our primary focus. We stand for freedom of expression and people being able to see all sides of any topic. That's put in jeopardy when abuse and harassment stifle and silence those voices. We won't tolerate it and we're launching new efforts to stop it.[68]

There are many more examples. Microsoft President Brad Smith in 2018 opined that "we should work to foster a healthier online environment more broadly. . . . [D]igital discourse is sometimes increasingly toxic. There are too many days when online commentary brings out the worst in people."[69] And testifying before Congress in 2018, Facebook CEO

[63] See, e.g., Objectionable Content, Community Standards, Facebook, https://www.facebook.com/communitystandards/objectionable_content [https://perma.cc/9TMH-R2HG] (last visited July 30, 2019).

[64] Josh Halliday, Twitter's Tony Wang: 'We Are the Free Speech Wing of the Free Speech Party,' Guardian (Mar. 22, 2012), https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech [https://perma.cc/75Z2-NBZP].

[65] Somini Sengupta, Twitter's Free Speech Defender, N.Y. Times (Sept. 2, 2012), https://www.nytimes.com/2012/09/03/technology/twitter-chief-lawyer-alexander-macgillivray-defender-free-speech.html (on file with the *Columbia Law Review*).

[66] See Nicholas Thompson, Jack Dorsey on Twitter's Role in Free Speech and Filter Bubbles, WIRED (Oct. 16, 2018), https://www.wired.com/story/jack-dorsey-twitters-role-free-speech-filter-bubbles/ [https://perma.cc/D6HJ-HJTQ].

[67] See id.

[68] Ed Ho, An Update on Safety, Twitter: Blog (Feb. 7, 2017), https://blog.twitter.com/en_us/topics/product/2017/an-update-on-safety.html [https://perma.cc/PA9C-HET6].

[69] Brad Smith, A Tragedy that Calls for More than Words: The Need for the Tech Sector to Learn and Act After Events in New Zealand, Microsoft (Mar. 24, 2019), https://blogs.microsoft.com/on-the-issues/2019/03/24/a-tragedy-that-calls-for-more-than-words-the-need-for-the-tech-sector-to-learn-and-act-after-events-in-new-zealand/ [https://perma.cc/ML64-JQTF].

Mark Zuckerberg concisely explained Facebook's shift in thinking this way: "It's not enough to just connect people. We have to make sure that those connections are positive. It's not enough to just give people a voice. We need to make sure that people aren't using it to harm other people or to spread misinformation."[70]

There are many more examples, but the point is that the major platforms now aspire to effective speech control to protect the "health" or "safety" of their platforms. But how to they do it? That is the subject of the next section.

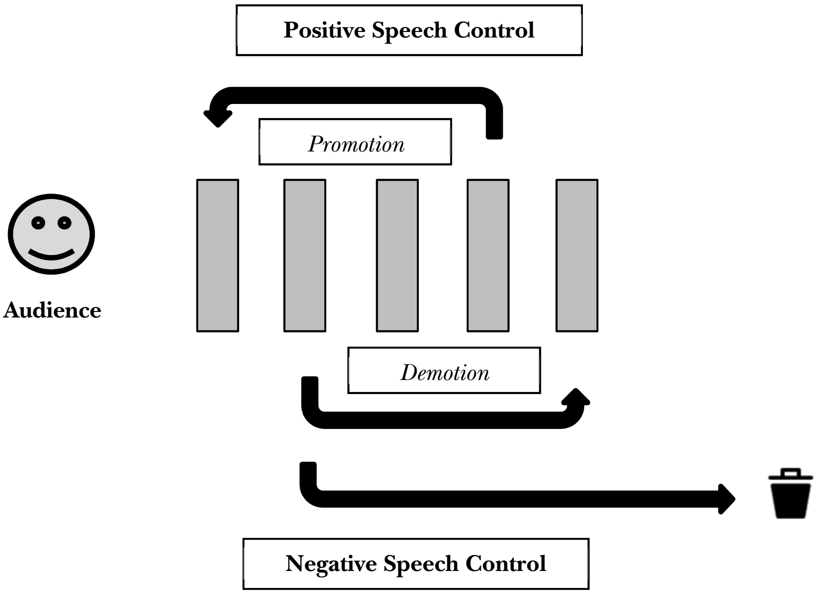
B. *How Platforms Control Speech*

The control of speech in the United States and the world is possibly the most advanced example of a hybrid human–machine system of social ordering that has replaced what was once primarily governed by law. All of the major speech platforms use a mixture of software, humans following rules, and humans deliberating to enforce and improve their content rules.[71]

[70] Transcript of Mark Zuckerberg's Senate Hearing, Wash. Post (Apr. 10, 2018), https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/ (on file with the *Columbia Law Review*).

[71] For recent articles offering a deeper investigation into how these platforms are shaping their content rules, see Kate Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, 131 Harv. L. Rev. 1598 (2018), and Simon van Zuylen-Wood, "Men Are Scum": Inside Facebook's War on Hate Speech, Vanity Fair (Feb. 26, 2019), https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech [https://perma.cc/3D7P-773L]. The author also attended a two-day Facebook seminar on its speech-control systems, from which some of this information is drawn.

FIGURE 1: HOW PLATFORMS CONTROL SPEECH



Speech is controlled by both affirmative and negative tools (promotion and suppression). *Affirmative* speech control entails choosing what is brought to the attention of the user. It is found in the operation of search results, newsfeeds, advertisements, and other forms of promotion and is typically algorithmic.[72] *Negative* speech control consists of removing and taking down disfavored, illegal, or banned content, and punishing or removing users.[73] The latter form of control, inherently more controversial, may be achieved in response to complaints, or proactively, by screening posted content.

Both positive and negative speech control have both human and algorithmic elements. Google's search results, the Facebook newsfeed, and the order in which tweets appear to Twitter users are all decided by algorithm.[74] In recent years, platforms like Facebook and Twitter have

[72] See, e.g., How Search Algorithms Work, Google, https://www.google.com/search/howsearchworks/algorithms/ [https://perma.cc/RY8V-58ZQ] [hereinafter Search Algorithms] (last visited July 31, 2019).

[73] Miguel Helft, Facebook Wrestles with Free Speech and Civility, N.Y. Times (Dec. 12, 2010), https://www.nytimes.com/2010/12/13/technology/13facebook.html?_r=0 (on file with the *Columbia Law Review*).

[74] See, e.g., Search Algorithms, supra note 72; Nicolas Koumchatzky & Anton Andryeyev, Using Deep Learning at Scale in Twitter's Timelines, Twitter: Blog (May 9, 2017), https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines.html [https://perma.cc/E464-DJKU]; Ramya Sethuraman, Jordi Vallmitjana & Jon Levin, Using Surveys to Make News Feed More Personal, Facebook

begun using their affirmative powers of promotion to disadvantage disfavored speech, by ranking it as "lower quality."[75] For example, facing criticism that it had aided the dissemination of fake news and propaganda, Facebook in 2017 announced it was reworking its algorithm to disfavor, among other things, posts that were untruthful.[76] And as part of its technical attack on abusive speech, Twitter suggested that its search results would avoid content algorithmically deemed abusive or of low quality.[77]

The negative methods of speech control on platforms—takedowns—were originally complaint driven and executed by humans.[78] On the major platforms, the takedowns were first implemented for nudity and pornography. Platforms like Facebook and YouTube kept pornography off of their platforms by employing humans to swiftly respond to complaints and took down almost all nudity or pornographic films.[79] Today, those systems have matured into large "content review systems" that combine human and machine elements.

Facebook has been the most transparent about its system. The human part is some 15,000 reviewers, most of whom are private contractors working at call centers around the world, coupled with a team of technical and legal experts based in Facebook's headquarters.[80]

Newsroom (May 16, 2019), https://newsroom.fb.com/news/2019/05/more-personalized-experiences/ [https://perma.cc/U9JL-QVUV].

[75] See, e.g., Varun Kacholia, News Feed FYI: Showing More High Quality Content, Facebook Bus. (Aug. 23, 2013), https://www.facebook.com/business/news/News-Feed-FYI-Showing-More-High-Quality-Content [https://perma.cc/422G-Z4UJ] (stating that Facebook's machine-learning algorithm counts reports that a post is "low quality" in deciding what content to show).

[76] Adam Mosseri, Working to Stop Misinformation and False News, Facebook for Media (Apr. 7, 2017), https://www.facebook.com/facebookmedia/blog/working-to-stop-misinformation-and-false-news [https://perma.cc/7D9L-8GKQ].

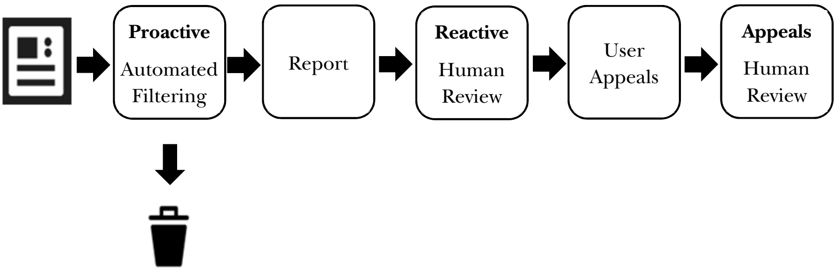
[77] See Donald Hicks & David Gasca, A Healthier Twitter: Progress and More to Do, Twitter: Blog (Apr. 16, 2019), https://blog.twitter.com/en_us/topics/company/2019/health-update.html [https://perma.cc/SRF7-Q2UR].

[78] See id. (stating that Twitter previously relied on reports to find abusive tweets).

[79] See Nick Summers, Facebook's 'Porn Cops' Are Key to Its Growth, Newsweek (Apr. 30, 2009), https://www.newsweek.com/facebooks-porn-cops-are-key-its-growth-77055 [https://perma.cc/5HRA-UMMM] (describing the job of Facebook's content moderators and the scope of its review system); Catherine Buni & Soraya Chemaly, The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech, The Verge (Apr. 13, 2016), https://www.theverge.com/2016/4/13/11387934/internet-moderator-hstory-YouTube-facebook-reddit-censorship-free-speech [https://perma.cc/U8PS-H6J8] (describing the job of content moderators in reviewing posts); Jason Koebler & Joseph Cox, The Impossible Job: Inside Facebook's Struggle to Moderate Two Million People, Vice (Aug. 23, 2018), https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works [https://perma.cc/4VFY-5FZ5] (describing the history of Facebook's content moderation system).

[80] See van Zuylen-Wood, supra note 71; Casey Newton, The Trauma Floor: The Secret Lives of Facebook Moderators in America, The Verge (Feb. 25, 2019), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona/ [https://perma.cc/F78T-AJY3].

FIGURE 2: CONTENT REVIEW AT FACEBOOK



In this system, forbidden content is flagged, and then sent to a human for review. If the human decides it violates the content guidelines, they take it down, and a notice is sent to the poster, who may ask for an appeal. The appeal is decided by a human; in hard cases, the appeal may go through several levels of review.[81]

In recent years, Facebook and the rest of the platforms have deployed intelligent software as an aid to this process. The first lines of defense are proactive filters which prevent certain forms of content from being posted at all. Among the first AI-driven negative speech controls was YouTube's Content ID system, first launched in 2007.[82] Content ID is software that compares uploaded videos against a database of copyrighted materials to determine whether the video is presumptively infringing a copyright.[83] If so, the copyright owner is automatically notified and given the choice of ordering the video taken down, or accepting a revenue-sharing agreement for any advertising revenue the video generates.[84] Since 2013 or so, the major platforms have used a similar system, PhotoDNA, that proactively detects videos of child pornography

[81] See van Zuylen-Wood, supra note 71 (describing the development of Facebook's appeals process).

[82] See Google, How Google Fights Piracy 24 (2018), https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Piracy_2018.pdf [https://perma.cc/9VHW-NX35] [hereinafter How Google Fights Piracy]; see also Sam Gutelle, The Long, Checkered History of YouTube's Attempt to Launch a Music Service, Tubefilter, (May 22, 2018), https://www.tubefilter.com/2018/05/22/youtube-music-service-history/ [https://perma.cc/D82U-P6HC] (describing the "wild west" history of the early days of music video distribution on YouTube prior to the launch of Content ID).

[83] See YouTube Help, How Content ID Works, supra note 39.

[84] See id. According to Google, the arrangement has yielded payments of over $3 billion for rights holders. How Google Fights Piracy, supra note 82, at 25.

and prevents them from being posted.[85] The major platforms have also installed proactive screens to block terrorist propaganda.[86] Hence, in testimony before Congress, Zuckerberg stated that "99 percent of the ISIS and Al Qaida content that we take down on Facebook, our A.I. systems flag before any human sees it."[87]

Proactively flagging hate speech and other forms of offensive speech is inherently more subjective than flagging nudity, copyright infringement, or child pornography. Nonetheless, Twitter and Facebook have begun using software to flag or take down such materials.[88] At Twitter in 2017, Dorsey pledged "a completely new approach to abuse" involving more proactive use of AI.[89] Twitter redesigned its search engine to create the option of hiding abusive content;[90] the platform also began systematically downgrading "low-quality" tweets.[91]

But what about the hard cases? In 2018, Facebook announced that it was planning to supplement its current review process with review conducted by a review board, acting in a court-like fashion, comprised of

[85] See Riva Richmond, Facebook's New Way to Combat Child Pornography, N.Y. Times: Gadgetwise (May 19, 2011), https://gadgetwise.blogs.nytimes.com/2011/05/19/facebook-to-combat-child-porn-using-microsofts-technology/ (on file with the *Columbia Law Review*) (reporting Facebook's adoption of PhotoDNA technology); Jennifer Langston, How PhotoDNA for Video Is Being Used to Fight Online Child Exploitation, Microsoft (Sept. 12, 2018), https://news.microsoft.com/on-the-issues/2018/09/12/how-photodna-for-video-is-being-used-to-fight-online-child-exploitation/ [https://perma.cc/2LKS-RBMN].

[86] See Klonick, supra note 71, at 1651–52 (describing how Facebook, YouTube, and Twitter came to monitor and remove terrorist content at the request of the government, then later on their own); Joseph Menn & Dustin Volz, Google, Facebook Quietly Move Toward Automatic Blocking of Extremist Videos, Reuters (June 24, 2016), https://www.reuters.com/article/internet-extremism-video/rpt-google-facebook-quietly-move-toward-automatic-blocking-of-extremist-videos-idUSL1N19H00I [https://perma.cc/25XD-9D9N].

[87] Transcript of Mark Zuckerberg's Senate Hearing, supra note 70.

[88] See Daniel Terdiman, Here's How Facebook Uses AI to Detect Many Kinds of Bad Content, Fast Company (May 2, 2018), https://www.fastcompany.com/40566786/heres-how-facebook-uses-ai-to-detect-many-kinds-of-bad-content [https://perma.cc/N924-NACX] (reporting on the details of Facebook's AI content-flagging system); Queenie Wong, Twitter Gets More Proactive About Combating Abuse, CNET (Apr. 17, 2019), https://www.cnet.com/news/twitter-gets-more-proactive-about-combating-abuse/ (on file with the *Columbia Law Review*) (noting Twitter claims that thirty-eight percent of all content that violates its terms of service is flagged automatically before a user reports it).

[89] Jack Dorsey (@jack), Twitter (Jan. 30, 2017), https://twitter.com/jack/status/826231794815037442 [https://perma.cc/7HGX-YV49]; see also Kurt Wagner, Twitter Says It's Going to Start Pushing More Abusive Tweets Out of Sight, Vox: Recode (Feb. 7, 2017), https://www.vox.com/2017/2/7/14528084/twitter-abuse-safety-features-update [https://perma.cc/6GQD-JBVA].

[90] Wagner, supra note 89 (reporting on Twitter's "safe search" feature and its use of "machine learning technology (a.k.a. algorithms) to automatically hide certain responses," which the user cannot opt out of).

[91] Jane Wakefield, Twitter Rolls Out New Anti-Abuse Tools, BBC (Feb. 7, 2017), https://www.bbc.com/news/technology-38897393 [https://perma.cc/2TC2-7MS4].

external, disinterested parties.[92] It is to that and other adjudicative bodies to which we now turn.

C. *The Reinvention of Adjudication*

Speech control by its nature produces hard problems. Is the phrase "kill all men" a form of hate speech, a joke (in context), or a feminist term of art?[93] If the phrase is taken down as hate speech, should it be put back up on review? The answer, of course, is that "it depends." The emergence of such problems has driven the major platforms to develop one or another forms of adjudication for this kind of hard case—a reinvention of the court, so to speak.

Lon Fuller defined an adjudication as a decision in which the participant is offered the opportunity to put forth "reasoned arguments for a decision in his favor."[94] By that measure, we can date the history of adjudicative content control on major online platforms to at least the mid-2000s.[95] In 2008, Jeffrey Rosen documented an early content-related deliberation at Google. It centered on a demand from the Turkish government that YouTube remove videos that the government deemed offensive to the founder of modern Turkey, in violation of local law. Rosen described the deliberation as follows:

> [Nicole] Wong [a Google attorney] and her colleagues set out to determine which [videos] were, in fact, illegal in Turkey; which violated YouTube's terms of service prohibiting hate speech but allowing political speech; and which constituted expression that Google and YouTube would try to protect. There was a vigorous internal debate among Wong and her colleagues at the top of Google's legal pyramid. Andrew McLaughlin, Google's director of global public policy, took an aggressive civil-libertarian position, arguing that the company should pro-

[92] See Mark Zuckerberg, A Blueprint for Content Governance and Enforcement, Facebook (Nov. 15, 2018), https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/ [https://perma.cc/XU5H-Z2UV]; see also Ezra Klein, Mark Zuckerberg on Facebook's Hardest Year, and What Comes Next, Vox (Apr. 2, 2018), https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge [https://perma.cc/3N7X-AT2X] (noting Zuckerberg's early 2018 intention to form an independent appeal process for users to challenge Facebook's content-moderation decisions).

[93] A description of the debate over this phrase can be found at stavvers, Kill All Men, Another Angry Woman (May 7, 2013), https://anotherangrywoman.com/2013/05/07/kill-all-men/ [https://perma.cc/59ES-5EEU].

[94] Fuller, supra note 10, at 364.

[95] There were also earlier such speech controls on online platforms. For a very early case study of an adjudication and punishment in an online forum, see Julian Dibbell, A Rape in Cyberspace, Village Voice (Oct. 18, 2005), https://www.villagevoice.com/2005/10/18/a-rape-in-cyberspace/ [https://perma.cc/2F5R-JVHY], See also Klonick, supra note 71, at 1618–21 (summarizing YouTube's, Facebook's, and Twitter's differing early approaches to content moderation, all overseen by lawyers normatively influenced by First Amendment principles).

> tect as much speech as possible. Kent Walker, Google's general counsel, took a more pragmatic approach, expressing concern for the safety of the dozen or so employees at Google's Turkish office. The responsibility for balancing these and other competing concerns about the controversial content fell to Wong, whose colleagues jokingly call her "the Decider . . . ."[96]

Since the mid-2010s, the platforms have developed larger and more specialized teams to adjudicate these kinds of hard problems, usually associated with the general counsel's office, and labeled the "Trust and Safety Council" (Twitter) or "safety and security" (Facebook).[97] Twitter, like Facebook, faces questions that emerge from complaints of abuse and propagandizing by political figures.[98] Its mechanism is a highly deliberative policy group centered in the general counsel's office to address the hardest speech problems.[99] Within the policy group is a leadership council that constantly updates the content guidelines applied by its reviewers.[100] The leadership council, which includes CEO Jack Dorsey, acts, in effect, as Twitter's supreme speech-moderation authority, and is responsible for both tough cases and large changes in policy.[101] It was through the deliberations of this group that, for example, Twitter decided to create more tools for screening "dehumanizing speech" in September 2018.[102] Here is how Twitter described its "dehumanizing speech" policy, outlined in a document not unlike that of a government agency promulgating a new rule:

> Language that makes someone less than human can have repercussions off the service, including normalizing serious violence. Some of this content falls within our hateful conduct policy . . . but there are still Tweets many people consider to be abusive, even when they do not break our rules. Better addressing this gap is part of our work to serve a healthy public conversation.

[96] Jeffrey Rosen, Google's Gatekeepers, N.Y. Times Mag. (Nov. 28, 2008), https://www.nytimes.com/2008/11/30/magazine/30google-t.html (on file with the *Columbia Law Review*).

[97] See Sara Harrison, Twitter and Instagram Unveil New Ways to Combat Hate—Again, WIRED (July 11, 2019), https://www.wired.com/story/twitter-instagram-unveil-new-ways-combat-hate-again (on file with the *Columbia Law Review*).

[98] See Interview by Will Oremus with Vijaya Gadde, Gen. Counsel, Twitter (July 19, 2018), https://slate.com/technology/2018/07/twitters-vijaya-gadde-on-its-approach-to-free-speech-and-why-it-hasnt-banned-alex-jones.html [https://perma.cc/VPY8-SWM2] (quoting Twitter's general counsel as saying that "philosophically, [Twitter] ha[s] thought very hard about how to approach misinformation, and . . . felt that we should not as a company be in the position of verifying truth").

[99] See id.

[100] Telephone Interview with Vijaya Gadde, Legal, Policy, & Trust and Safety Lead, Twitter (Mar. 29, 2019).

[101] Id.

[102] See Vijaya Gadde & Del Harvey, Creating New Policies Together, Twitter: Blog (Sept. 25, 2018), https://blog.twitter.com/en_us/topics/company/2018/Creating-new-policies-together.html [https://perma.cc/W6TR-EJS9].

> With this change, we want to expand our hateful conduct policy to include content that dehumanizes others based on their membership in an identifiable group, even when the material does not include a direct target.[103]

We have already discussed Facebook's basic system of review.[104] Similar to Twitter, it currently has an internal policy group that works on hard cases and updates to policies in response to such cases.[105] To supplement and replace parts of the appeal process, the firm in 2018 announced plans to create an independent review board.[106] As Zuckerberg explained the idea,

> You can imagine some sort of structure, almost like a Supreme Court, that is made up of independent folks who don't work for Facebook, who ultimately make the final judgment call on what should be acceptable speech in a community that reflects the social norms and values of people all around the world.[107]

According to Facebook, the board would be independent, with approximately forty members, and sit in panels of three[108] to review "hard cases."[109] They would be brought the hardest questions arising from content control on Facebook, and release their written decisions in two weeks.[110] The panels would have the ability to overrule Facebook's decisions and make policy suggestions, but not to rewrite the content rules themselves.[111]

Here, in summary form, we have a sense of how machines and humans combine to control speech on the major online platforms. We can now address the question of whether this institutional framework offers any promise for the future.

III. Hybrid Systems and the Comparative Advantages of Software and Courts

This Part addresses the comparative advantages of software and courts, and offers a normative defense of hybrid systems.

[103] Id.

[104] See supra notes 75–81 and accompanying text.

[105] See Klein, supra note 92.

[106] See supra note 92 and accompanying text.

[107] Klein, supra note 92.

[108] Facebook, Draft Charter: An Oversight Board for Content Decisions 3 (2019), https://fbnewsroomus.files.wordpress.com/2019/01/draft-charter-oversight-board-for-content-decisions-2.pdf [https://perma.cc/E9ZX-EDVF] [hereinafter Draft Charter].

[109] Nick Clegg, Charting a Course for an Oversight Board for Content Decisions, Facebook Newsroom (Jan. 28, 2019), https://newsroom.fb.com/news/2019/01/oversight-board [https://perma.cc/3F6Q-EZ9X].

[110] Draft Charter, supra note 108, at 5.

[111] Id. at 3.

A. *Will Software Eat the Law?*

The case study of the online control of speech has shown the tendency of rule-based systems to generate hard and easy cases, giving rise to crude hybrid systems designed to manage that challenge. This Part seeks to theorize some of the advantages of hybrid systems. This returns us to the central question: Will software eat the law? (Or, as asked here, will software tools take over almost all of online content control?) In our case study, for routine matters, the answer is already yes, because of the undeniable comparative advantage of software in matters of scale, speed, and efficacy. To ask this question is a little like asking, a century ago, whether motorized lawnmowers might take over the mowing of lawns. But as to whether software will or should replace everything, the answer is no.

It is important to be more precise as to why this is so. As a means of regulation, software's main advantage over legal systems lies in what law would call its enforcement capacity.[112] Code is fast, can scale to meet the size of the problem, and operates at low marginal cost. But there is more to it than that. Code can be designed, as Lessig first pointed out, to change the very architecture of decision, the option set, and the menu of choices faced.[113] Consider that, when it comes to child pornography, the main platforms don't just ban it and punish transgressors but remove the option of posting it in the first place.[114] The enforcement mechanism is therefore ex ante rather than ex post, and hence vastly more effective than law, which always acts after a wrong is committed.

But if intelligent software is effective, it is also inherently inhuman, and prone, at least for the foreseeable future, to make absurd errors that can be funny, horrific, or both. Following rules blindly leads to consequences like the takedown of famous paintings as "nudity."[115] Software also faces limits of explainability, which is a problem for legal decision-making. Software can often explain *how* it reached a decision, but not *why*.[116] That may be fine for a thermostat, but is a limitation for a system that is supposed to both satisfy those subjected to it and prompt acceptance of an adverse ruling.

As it stands, the decisions to take down content by Facebook or Twitter are, to users, nearly a black box, which is acceptable for routine

[112] See Wu & Talley, supra note 7.

[113] See Lessig, supra note 12, at 121–25.

[114] Richmond, supra note 85.

[115] See, e.g., Kerry Allen, Facebook Bans Flemish Paintings Because of Nudity, BBC: News from Elsewhere (July 23, 2018), https://www.bbc.com/news/blogs-news-from-elsewhere-44925656 [https://perma.cc/JH5Z-CR9P].

[116] See Ashley Deeks, The Judicial Demand for Explainable Artificial Intelligence, 119 Colum. L. Rev. 1829, 1832–38 (2019); cf. Tom Simonite, Google's AI Guru Wants Computers to Think More Like Brains, WIRED (Dec. 12, 2018), https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/ [https://perma.cc/ECJ7-3P5Z].

decisions, but in borderline cases have already provoked anger and dissatisfaction.[117] As Richard M. Re and Alicia Solow-Niederman warn, software decision-systems can be "incomprehensible, data-based, alienating, and disillusioning."[118]

This is what the speech control case study helps make clear. If the only goal in speech control was taking down as much forbidden material as quickly as possible, mistakes be damned, the discussion would be over. But you don't need to be a First Amendment scholar to suggest that this would hardly amount to a satisfying or successful system of speech control, or one that the public would accept. The lines governing the forbidden from the provocative are always fuzzy, and building a healthy speech environment, at the risk of stating the obvious, is more than building the fastest takedown machine. In fact, the engineer's thirst for efficacy can obscure the fact that what Facebook and other platforms are building can also be described, without exaggeration, as among the most comprehensive censorship machines ever built.

Nor can we ignore the fact that what counts as acceptable speech for billions of people around the world is currently being decided by a relatively small group of private actors in Northern California. To suggest that this creates questions of legitimacy in the decision of matters of interest to the public in many countries seems almost too obvious to state. Hence, based on both public dissatisfaction *and* poor results, a purely software-based replacement is a bad aspiration.

That's why the platforms are already turning to specialized human adjudicators, as a supplement to the software systems, to offer answers to some of these problems.[119] Their advantages—really the advantages of courts more generally—lie in two areas.

The first is procedural fairness. A group of legal theorists, of which Tom Tyler is best known, has for decades suggested that the best justification for the court system lies in providing a sense of procedural fairness to participants.[120] The empirical studies conducted by Tyler and others suggest that when litigants feel they have a voice and are treated with respect, they tend to be more accepting of decisions, even adverse outcomes.[121] It seems unlikely, in the near future, that people with a grievance will be more satisfied with a software decision than a human decision on an important complaint. In the future, having a major decision be made by a human may become a basic indicium of fairness; it

[117] See, e.g., Sam Levin, Julia Carrie Wong & Luke Harding, Facebook Backs Down from 'Napalm Girl' Censorship and Reinstates Photo, Guardian (Sept. 9, 2016), https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo [https://perma.cc/78CB-KN32].

[118] Re & Solow-Niederman, supra note 2, at 242.

[119] See supra text accompanying notes 107–111.

[120] See Tom R. Tyler, Procedural Justice and the Courts, 44 Ct. Rev. 26, 30–31 (2007) (summarizing research in this area).

[121] See id. at 26.

is implicit in the emergence of what Aziz Z. Huq calls a "right to a human decision."[122]

That said, it is possible that our taste for human adjudication might be fleeting; perhaps it is akin to an old-fashioned taste for human travel agents. Eugene Volokh argues that any preference for human decision may turn out to be temporary, because humans are imperfect as well.[123] He believes that if an AI judge produces good decisions and good opinions, it will be broadly accepted, particularly if it is cheaper for users.[124] Volokh, characteristically, overstates his point, but he is right that there are in fact many areas where "impartial" code is trusted more than humans (compare Google Maps to asking for directions).[125] But that acceptance turns very heavily on the quality of decisions, to which we now turn.

The second benefit of human courts over software is their advantages in hard cases, and the prevention of absurd errors, obviously unjust results, and other inequitable consequences of a blind adherence to rules. There are, on closer examination, several ways in which a case can be "hard." Some cases might be hard only because the software lacks the ability to understand context or nuance, as in understanding that "I'm going to kill my husband" may be a figurative statement, not a death threat. And, others may be hard in the jurisprudential sense because they require the balancing of conflicting values or avoidance of absurd consequence. Finally, it may be that the stakes just seem large enough to merit human involvement, as in the decision to sentence someone to death. In each of these cases, the use of humans may prevent what Re and Solow-Niederman believe will be a tendency of AI systems to promote "codified justice at the expense of equitable justice."[126] How so? The premise is that leaving the hard cases to people of good character who are asked to listen to reasoned argument will have an effect, and that the effect will be positive for the rule system in question.

The theoretical support for this position is one of ancient pedigree and comes from the idea that something happens when intelligent, experienced, and thoughtful humans are asked to hear reasoned argument and the presentation of proofs to determine how a dispute should be settled. Over the centuries, the mental process accompanying the judicial process has been described in many different ways.[127] In the Anglo American tradition, it was presented in the semi-mystical idea that judges "discover" the law in the process of adjudication and deliberation, a law

[122] Huq, supra note 14, at 2.

[123] See Volokh, supra note 2, at 1170–71.

[124] Id.

[125] See also Tim Wu, The Bitcoin Boom: In Code We Trust, N.Y. Times (Dec. 18, 2017), https://www.nytimes.com/2017/12/18/opinion/bitcoin-boom-technology-trust.html (on file with the *Columbia Law Review*) (arguing that there are sometimes reasons to trust in code).

[126] Re & Solow-Niederman, supra note 2, at 255, 260.

[127] See id. at 252–53.

that was usually thought to be God given.[128] Blackstone writes of judges discovering the "the eternal, immutable laws of good and evil, to which the creator himself in all his dispensations conforms; and which he has enabled human reason to discover, so far as they are necessary for the conduct of human actions."[129]

Blackstone's theory that the law is best discovered by tuning into heavenly emanations enjoys a more limited following today.[130] But the idea that a particular mental process accompanies adjudication survives, even in the work of those highly critical of natural law reasoning. It is found in Llewellyn's idea of a judge's understanding of the "real rules" as distinct from the paper rules, and the skill involved in weighing demands of flexibility and stability in a legal system.[131] The judicial process is also a major part of Ronald Dworkin's theory of legal reasoning, which suggests that judges, when facing hard cases, begin to fill in gaps or conflicts through a process of rights-driven moral reasoning.[132] Hence, as Dworkin wrote in *Taking Rights Seriously*, a court won't let the son who murders his grandfather inherit wealth, not based on the following of any rule, but by reaching for the principle that doing so would be morally wrong.[133]

One does not need not to accept or agree with Dworkin's particular theory of how judges decide hard cases to accept that he has gotten at something important in the mechanism of decisionmaking. Richard Posner, for example accepts the premise that a judge, when deciding a hard case, exercises powers of intuitive judgment, though Posner believes they should be powers of pragmatic judgment.[134] Posner, who was a judge, wrote of the process this way: Judges necessarily "consider the implication of [their] interpretation for the public good" and, when making decisions about private rights, "consider the social consequences of alternative answers."[135] Or perhaps Fuller was correct when he asserted that the key is not labeling a person a judge, so much as the entire process of adjudication. He located the special sauce, such as it is, as "the presentation of proofs and reasoned arguments," yielding an expectation that the decision "meet the test of rationality."[136]

Cynics reading the preceding paragraphs might think that all that is being described is a bunch of hoodoo voodoo, a mystic secret sauce that

[128] 1 William Blackstone, Commentaries *40.

[129] Id.

[130] See, e.g., John S. Baker, Jr., Natural Law and Justice Thomas, 12 Regent U. L. Rev. 471, 471 (1999) (describing and defending the use of natural law approaches).

[131] See Frederick Schauer, Introduction to Llewellyn, *in* The Theory of Rules, supra note 17, at 11–13.

[132] See Dworkin, supra note 16, at 81–88.

[133] Id. at 23–28.

[134] See Richard A. Posner, Pragmatic Adjudication, 18 Cardozo L. Rev. 1, 5–8 (1996).

[135] Richard A. Posner, What Am I? A Potted Plant?: The Case Against Strict Constructionism, New Republic, Sept. 28, 1987, at 23, 23.

[136] See Fuller, supra note 10, at 365–70.

is hiding nothing more than judicial whim. Be that as it may, even such whims remain hard to replicate using artificial intelligence. And what Blackstone, Llewellyn, Dworkin, and Posner are all getting at is familiar to anyone who has either sat as a judge, or been asked to decide a hard case. The process brings forth a series of instincts, competing intuitions that can be of differing strengths in different people, but whose existence cannot be denied. A good account is given by Benjamin Cardozo, who, in *The Judicial Process*, describes a judge in deliberation as bombarded by competing forces, not all conscious.[137] Judges often ruminate at length, change their mind, want more facts, and want to consider different futures based on what they are proposing to do. Some may secretly (or openly, like Blackstone) believe that they are tapping into the divine, or, for those who claim a more secular mindset, the immutable principles of moral philosophy.

That said, returning to this Essay's case study and our times, it must be admitted that hoping for a Herculean process of judicial reasoning may be expecting a lot from the first hybrid systems, like the Facebook review board and its part-time judges. The court will have many disadvantages, including a lack of history, lack of traditions, lack of connection with government, and smaller matters like the probable lack of a courtroom (though perhaps robes will be provided). Fuller's idea that the setting and context matter may be right, and if so the Facebook appeals board may never inspire the kind of reasoning that garners respect.

In contrast, while I doubt it, it is possible that AI systems will soon begin to replicate the adjudicatory function in a manner indistinguishable from a human, while becoming able to explain what they are doing in a manner that complainants find acceptable.[138] And that, perhaps, will inspire people to trust such programs as less fallible than humans. Then the question will be whether judges are more like travel agents or more like spouses—whether being human is essential to the role. But for the foreseeable future, there is nothing that has anything close to these abilities; what we have is software intelligent enough to follow rules and replicate existing patterns, but that's about it. That's what makes hybrid systems seem almost inevitable, at least should we want social ordering to have any regard for the demands of justice, equity, or other human values.

B. *Implications and Other Counterarguments*

Reflecting their roots as software companies, the leaders of Silicon Valley firms usually state their ambition to have intelligent software even-

[137] See Cardozo, supra note 18, at 10–12.

[138] See Louise A. Dennis & Michael Fisher, Practical Challenges in Explicit Ethical Machine Reasoning, ArXiv (Jan. 4, 2018), https://arxiv.org/pdf/1801.01422.pdf [https://perma.cc/69UG-G3FK] (reviewing the several practical challenges AI systems face in replicating ethical reasoning).

tually solve problems by replacing humans entirely.[139] For example, self-driving cars are not designed as aids to driving, but as replacements for human drivers.[140] The industry has expressed similar goals for content-control systems, as in Facebook's promise to Congress that control of hate speech will be automated in the next five to ten years.[141]

This is the wrong aspiration. While the desire to have more effective and efficient systems of social control is understandable, far too much would be lost. The implication of this Essay is that the designers of intelligent software produced for social ordering should be aiming for the autopilot, not the self-driving car. The reasons why have already been stated; but until a computer is able to replicate not only a judge but the entire process of adjudication, we remain far short of an AI solution.

Similarly, as government begins to automate parts of the legal system (as has happened in limited ways already), a hybrid system should be the aspiration as well. Routine matters, like routine motion practice, and even perhaps frivolous cases, might well be automated to reduce the workload of the judiciary. The promise of doing so is not just saving costs but giving the judiciary more room to emphasize justice in the individual case as it devotes less of its time to reducing the judicial workload. Since the 1980s, numerous critics have pointed out that the huge increases in federal court filings have created a workload crisis.[142] As Judge Roger Miner wrote in 1997, "The situation has been deteriorating for many years and, although the courts have been attempting to cope by using various methods to accommodate the growing caseload traffic, the problems associated with volume largely remain unresolved."[143] One reaction has been the creation of various judicial doctrines designed to cope with the workload, from easier standards of dismissals, various means of reducing jurisdiction, plea bargaining in criminal cases, and reduced oral arguments.[144] With the help of software to handle routine procedural matters and even the decision of routine motions, government courts and judges might be able to devote more time and effort to the hard cases and im-

[139] See, e.g., Kevin Roose, A Machine May Not Take Your Job, but One Could Become Your Boss, N.Y. Times: The Shift (June 23, 2019), https://www.nytimes.com/2019/06/23/technology/artificial-intelligence-ai-workplace.html (on file with the *Columbia Law Review*).

[140] See Jonathan Vanian, Will Replacing Human Drivers with Self-Driving Cars Be Safer?, Fortune (June 14, 2017), http://fortune.com/2017/06/14/ford-argo-ai-self-driving-cars/ [https://perma.cc/7KWH-YAGU] ("[A]ccording to Bryan Salesky, the CEO of [Ford Motor Company's subsidiary] Argo AI, . . . [t]he rise of self-driving cars will usher a 'much safer mode of transportation' by 'removing the human from the loop' . . . .").

[141] Transcript of Mark Zuckerberg's Senate Hearing, supra note 70.

[142] See Cara Bayles, Crisis to Catastrophe: As Judicial Ranks Stagnate, 'Desperation' Hits the Bench, Law360 (Mar. 19, 2019), https://www.law360.com/articles/1140100 [https://perma.cc/YW3D-3ULZ]; see also Roger J. Miner, Book Review, 46 Cath. U. L. Rev. 1189, 1189–91 (1997) (reviewing Richard A. Posner, The Federal Courts: Challenge and Reform (1996) [hereinafter Posner, The Federal Courts]).

[143] Miner, supra note 142, at 1189.

[144] See Posner, The Federal Courts, supra note 142, at 160–85.

provement of the rules without the need to be constantly concerned about the impact of their decisions on their own workload.

This Essay could be wrong either on descriptive or normative grounds. Descriptively, it could turn out to be wrong that human judges have any lasting advantages over software; if an AI can pass a Turing test, it might well soon begin to replicate that which we call justice, and people could get used to decisions made by a machine. Or the opposite could be true: AI has often been grossly overrated and software might not make the inroads expected, leaving the legal system and other systems of social ordering more or less intact. There is no real way to address either of these objections other than to say that prediction is hard, especially when it comes to the future.

Normatively, it could also be wrong to think that there is really anything appealing about a hybrid human–machine system. Anthropologists like Hugh Gusterson write about the rise of the "roboprocess"—systems, like the U.S. credit rating system, that combine software with humans but actually disempower and deskill the humans employed by them.[145] Re and Solow-Niederman argue that introducing more software into the justice system will drive a shift in norms toward "codified" (that is, rule-driven) justice, as opposed to equitable justice.[146] They are not optimistic about adding humans, believing that "[r]etaining a human in the system . . . could succeed in preserving the legal system's preexisting public legitimacy—but only by objectionably sacrificing efficiency and uniformity that pure AI adjudication would otherwise offer."[147] The worst version of the hybrid system would pair the unthinking brutality of software-based justice with a token human presence designed to appease the humans subject to it. Such a system might arise out of cost cutting, in the manner that automated assistants are used in customer support to save money rather than improve service. This argument does make clear the danger of judging the judicial system by its costs alone, when the stakes are so much higher.

This suggests that the key question is the human–machine interface in a hybrid system. Just when and why are decisions brought to human attention, and who decides when a human should decide? Stated differently, how do we distinguish between "easy" and "hard" questions? It quickly becomes apparent that the human cases must include not just those that are hard in a jurisprudential sense, but also those where the stakes are large. The automated dispenser of speeding tickets may be one thing, but it is hard to imagine the fully automated assignment of the

[145] See Hugh Gusterson, Introduction: Robohumans, *in* Life by Algorithms, How Roboprocesses Are Remaking Our World, 1, 13–26 (Catherine Besteman & Hugh Gusterson, eds. 2019).

[146] See Re & Solow-Niederman, supra note 2, at 246–47.

[147] Id. at 284–85.

death sentence, even if it were shown, as compared to a jury, to more reliably determine guilt or innocence.[148]

Deciding what and when questions go to an empowered human is difficult out of context, but the most obvious model is a certiorari system used by appellate courts—a human system designed to decide when to decide. Whether that system is itself human or machine-run, or another hybrid makes for an interesting design problem. In any event, setting the border between human and machine decision is surely the linchpin of a successful hybrid system.

It might also be that hybrid systems accelerate a privatization of public justice. For some decades, with the rise of measures like compulsory arbitration, critics have complained that American justice has been privatized, usually in a manner designed to disfavor consumers, patients, and other weak groups.[149] The hybrid systems in the case study are all private adjudicators and policymakers. Their speech codes are created in-house, without traditional forms of public input. If successful, they may become a model whereby more and more areas of social ordering become subjects of such private hybrid systems.

It would be foolish to ignore such concerns. The topic of speech control may be a special case, given that the Supreme Court has effectively privatized speech control with its aggressive interpretations of the First Amendment.[150] But if we consider privatization of justice, the right answer might be "if you can't beat 'em, join 'em": The increased use of software may help improve the efficiency of routine justice, protecting the resources of the court system for preventing error in cases of either greater consequence or greater difficulty. Robot courts are not the right aspiration, but an augmented equivalent may very well be.

CONCLUSION

The comparative advantages of human, machine, and cyborg systems have been a longstanding subject of science fiction. But as the science fiction slowly becomes reality, one of the genre's longstanding predictions is coming true. It takes great effort to preserve human values when new technologies make it so easy to maximize efficient operations. There are reasons beyond the literary that so much science fiction is dystopian.

[148] When it comes to war, a parallel debate concerns the deployment of autonomous weapons. See generally Amanda Sharkey, Autonomous Weapons Systems, Killer Robots and Human Dignity, 21 Ethics & Info. Tech. 75 (2019) (exploring criticisms of autonomous weapon systems as violating human dignity).

[149] See, e.g., Jessica Silver-Greenberg & Michael Corkery, In Arbitration, a 'Privatization of the Justice System,' N.Y. Times: Dealbook (Nov. 1, 2015), https://www.nytimes.com/2015/11/02/business/dealbook/in-arbitration-a-privatization-of-the-justice-system.html (on file with the *Columbia Law Review*).

[150] See, e.g., Reno v. ACLU, 521 U.S. 844, 870 (1997) (striking down the Communications Decency Act and holding that the internet is due the highest level of First Amendment protection).

Columbia Law Review. ISSN 0010-1958.

Published by Columbia Law School students monthly except for February, July, August, and September. Printed in Hanover, Pennsylvania. Editorial Office, 435 West 116th Street, New York, New York 10027. Subscription per volume $60.00 domestic, $85.00 foreign, single issues $20.00 domestic, $30.00 foreign. Prices for special/symposium issues may vary. These figures include postage. Prepayment requested on all orders. If a subscription is to be discontinued at expiration, notice to that effect should be sent; otherwise it will be renewed as usual. For issues prior to volume 105 inquire of William S. Hein & Co., Inc., 1285 Main Street, Buffalo, New York 14209-1987, 1-800-828-7571, http://www.wshein.com. Back issues are also available in PDF format through Hein Online, http://www.heinonline.org.

Change of Address: Send your change of address to the *Review* at 435 West 116th Street, New York, New York 10027, at least 30 days before the date of the issue with which it is to take effect. The Post Office will not forward copies unless you provide extra postage; duplicate copies will not be sent. POSTMASTER: Send Address Changes to Columbia Law Review, 435 West 116th Street, New York, New York 10027. Periodicals postage paid at New York, New York, and at additional mailing offices.

Information for Contributors: Unsolicited manuscripts will be considered for publication. All manuscripts should be double spaced. Text and citations should conform to *The Bluebook: A Uniform System of Citation* (20th ed. 2015). As a matter of policy, the *Review* encourages the use of gender-neutral language. The *Review* requires contributors to submit pieces electronically via the instructions on its Submissions page. Please visit http://www.columbialawreview.org/submissions for further information regarding submissions.

Internet Address: The *Columbia Law Review* homepage is located at http://www.columbialawreview.org.

VOL. 119, NO. 7

COLUMBIA LAW REVIEW

Pages
1773
to
2028

NOVEMBER 2019