34:3 BERKELEY TECHNOLOGY LAW JOURNAL

2019

Pages

705

to

918

Berkeley Technology Law Journal

Volume 34, Number 3

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

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

Berkeley Technology Law Journal
University of California
School of Law
3 Law Building
Berkeley, California 94720-7200
editor@btlj.org
https://www.btlj.org

BERKELEY TECHNOLOGY LAW JOURNAL

VOLUME 34 NUMBER 3 2019

TABLE OF CONTENTS

ARTICLES

SUBSCRIBER INFORMATION

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

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

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

Form. The text and citations in the *Journal* conform generally to the THE CHICAGO MANUAL OF STYLE (16th ed. 2010) and to THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (Columbia Law Review Ass'n et al. eds., 20th ed. 2015). Please cite this issue of the *Berkeley Technology Law Journal* as 34 BERKELEY TECH. L.J. ___ (2019).

BTLJ ONLINE

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

INFORMATION FOR AUTHORS

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

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

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

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

DONORS

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

Partners

COOLEY LLP

FENWICK & WEST LLP

HOGAN LOVELLS

ORRICK, HERRINGTON & SUTCLIFFE LLP

WHITE & CASE LLP

Benefactors

BAKER BOTTS LLP

COVINGTON & BURLING LLP

FISH & RICHARDSON P.C.

JONES DAY

KIRKLAND & ELLIS LLP

LATHAM & WATKINS LLP

MCDERMOTT WILL & EMERY

MORRISON & FOERSTER LLP

POLSINELLI LLP

SIDLEY AUSTIN LLP

WEIL, GOTSHAL & MANGES LLP

WILMER CUTLER PICKERING HALE AND DORR LLP

WILSON SONSINI GOODRICH & ROSATI

WINSTON & STRAWN LLP

Corporate Benefactors

ATLASSIAN

COMPUTER & COMMUNICATIONS INDUSTRY ASSOCIATION

CORNERSTONE RESEARCH

FUTURE OF PRIVACY FORUM

GOOGLE, INC.

HEWLETT FOUNDATION, THROUGH THE CENTER FOR LONG-TERM CYBERSECURITY

INTEL

INVENTIONSHARE

LITINOMICS

MICROSOFT CORPORATION

MOZILLA

NERA ECONOMIC CONSULTING

NOKIA

PALANTIR

RLM TRIALGRAPHIX

THE WALT DISNEY COMPANY

Members

Baker & McKenzie LLP

Crowell & Moring

Desmarais LLP

Durie Tangri LLP

Finnegan, Henderson, Farabow, Garrett & Dunner, LLP

GTC Law Group LLP & Affiliates

Haynes and Boone, LLP

Hickman Palermo Becker Bingham

Irell & Manella LLP

Keker Van Nest & Peters LLP

Kilburn & Strode

Kilpatrick Townsend & Stockton LLP

Knobbe Martens LLP

Kwan & Olynick LLP

Morgan, Lewis & Bockius LLP

Paul Hastings LLP

Robins Kaplan LLP

Ropes & Gray LLP

Simpson Thacher & Bartlett LLP

Troutman Sanders LLP

Van Pelt, Yi & James LLP

Weaver Austin Villeneuve & Sampson LLP

BOARD OF EDITORS 2018–2019

Executive Board

Editor-in-Chief
ELLE XUEMENG WANG

Managing Editor
CHRISTOPHER BROWN

Senior Articles Editors
LAURA KELLEY
ALEX BARATA
YARDEN KAKON

Senior Executive Editor
ANGELA LYONS-JUSTUS

Senior Production Editor
MEGAN MCKNELLY

Senior Scholarship Editors
AMIT ELAZARI
NOMI CONWAY

Senior Annual Review Editors
CHRIS CHUANG
KATHARINE CUMMINGS

Senior Online Content Editor
KATHERINE BURKHART

Editorial Board

Submissions Editors
DANIEL CHASE
BEATRICE NYBERT

Production Editors
CHELSEY MORI
HAILEY YOOK
JANELLE LAMB
LOUISE DECOPPET

Technical Editors
MARIA BELTRAN
COLIN RAVELLE

Annual Review Editors
JULEA LIPIZ
AISLINN SMALLING

Notes & Comments Editors
NICK CALCATERRA
RYAN KWOCK

Symposium Editors
DARIUS DEHGHAN
JESSICA HOLLIS

Web Content Editors
CONCORD CHUNG
CHANTE ELIASZADEH

Podcast Editor
MIRANDA RUTHERFORD

LLM Editors
CRISTINA DE LA PAZ
MAYARA RUSKI
AUGUSTO SA

Member Relations Editor
CLARA CHOI

Alumni Relations Editor
NIR MAOZ

External Relations Editor
ANDREA SHEN

Articles Editors

CHELSEA ANDRE
TIAGO AQUINO
SAVANNAH CARNES
ELSIE CHEANG
SHERILYN CHEW

NITESH DARYANNANI
LESLIE DIAZ
GRACE FERNANDEZ
VESTA GOSHTASBI
ALEXANDER KROIS
NINA MILOSAVLJEVIC

CRISTINA MORA
COURTNEY REED
WESLEY TIU
MEI XUAN
LI ZHANG

MEMBERSHIP

Vol. 34 No. 3

Associate Editors

MADISON BOWER
MUHTADI CHOUDHURY
MATTHEW CHUNG
ELLA PADON CORREN
CHRISTINA CROWLEY
DAVID FANG
YUAN FANG
HARRISON GERON
ADRIAN KINSELLA
CLARA KNAPP
VICKY WEI-CHI LEE
HUI-FANG LIN
ALLAA MAGEID
ARYEH PRICE
MARTA STUDNICKA
THERESA TAN
EMILEE WU
CHENZHAO YU
MICHELLE ZIPERSTEIN

Members

GABRIELA ABREU
RIDDHI ADHIKARI
SAFIYA AHMED
MIMANSA AMBASTHA
NISHANT ANURAG
ARTIN AU-YEUNG
EMILY AVAZIAN
CHRISTOPHER BARCLAY
VERONICA BOGNOT
ALEXIS CALIGIURI
ANNA ESCRIGAS CANAMERAS
LAUREN CARROLL
CLAIRE CHANG
ALEX CHEMERINSKY
KEVIN CHEN
TIFFANY CHEN
YITING CHENG
ARI CHIVUKULA
CLAIRE CHRISTENSEN
SCOT CONNER
JULIEN CROCKETT
ARPITA DAS
SAMAPIKA DASH
KERENSA GIMRE
KELLY GO
SARAH GOLD
ANDREW GORIN
KEVIN GU
PETER GUTMAN
DEREK HA
CATHERINE HARRIS
DYLAN HELGESON
ERIN HILLIARD
ALLAN HOLDER
JEONGHOON HONG
FIONA HUANG
VICTORIA CONSTANCE HUANG
JONATHAN HUANG
MAISIE IDE
YING JIANG
SAACHI JUNEJA
GIA JUNG
MARGARETH KANG
CHAITANYA KAUSHIK
JORGE KINA
ERIN MOORE
WALTER MOSTOWY
LIU QING NG
YUTA OISHI
BIHTER OZEDIRNE
VALINI PANTA
SORA PARK
CHULHYUN PARK
AYESHA RASHEED
NOELLE REYES
BRYCE ROSENBOWER
JOHN RUNKEL
RYAN JORGENSEN
ARMBIEN SABILLO
GINETTA SAGAN
TARINI SAHAI
SHREYA SANTHANAM
JOSH SEDGWICK
EVAN SEEDER
VICKY EL KHOURY SFEIR
YAAMINI SHARMA
CARMEN SOBCZAK

Trenton Davis
Liz Douglass
Ida Ebeid
Rucha Ekbote
Katelyn Feliciano
Olgamaris Fernandez
Moritz Flechsenhar
Yesenia Flores
Jason Francis
Logan Freeborn
Ravin Galgotia
Maribel Garcia
Garima Garg

Danny Koningisor
Michael Kostukovsky
Kristina Krasnikova
Soumya Jogaiah Krishnaraju
Emma Lee
Jian Lee
Killian Lefevre
Elly Leggatt
Xiaocao Li
Ashleigh Lussenden
Aartika Maniktala
Marissa Medansky
Alex Milne
Alexandre Mochon

Schuyler Standley
Lyric Stephenson
Daniel Todd
Viviane Trojan
Daniel Twomey
Edgar Vega
Yuhan Wang
Grace Winschel
Jolene Xie
Kevin Yang
Alison Yardley
Clark Zhang
Jieyu Zhang
Pengpeng Zhang
Evan Zimmerman

BTLJ ADVISORY BOARD

BERKELEY CENTER FOR LAW & TECHNOLOGY 2018–2019

Executive Director

JIM DEMPSEY

Faculty Directors

KENNETH A. BAMBERGER	PETER S. MENELL	PAMELA SAMUELSON
CATHERINE CRUMP	ROBERT P. MERGES	PAUL SCHWARTZ
CATHERINE FISK	DEIRDRE K. MULLIGAN	ERIK STALLMAN
CHRIS HOOFNAGLE	TEJAS NARECHANIA	JENNIFER M. URBAN
SONIA KATYAL	ANDREA ROTH	MOLLY S. VAN HOUWELING

Fellow

KATHRYN HASHIMOTO

Staff Directors

JANN DUDLEY	IRYS SCHENKER
RICHARD FISK	MATTHEW RAY

THE INSTITUTIONAL LIFE OF ALGORITHMIC RISK ASSESSMENT

Alicia Solow-Niederman,[†] *YooJung Choi*[††] *& Guy Van den Broeck*[†††]

ABSTRACT

As states nationwide turn to risk assessment algorithms as tools for criminal justice reform, scholars and civil society actors alike are increasingly warning that this technological turn comes with complications. Research to date tends to focus on fairness, accountability, and transparency within algorithmic tools. Although attention to whether these instruments are fair or biased is normatively essential, this Article contends that this inquiry cannot be the whole conversation. Looking at issues such as fairness or bias in a tool in isolation elides vital bigger-picture considerations about the institutions and political systems within which tools are developed and deployed. Using California's Money Bail Reform Act of 2017 (SB 10) as an example, this Article analyzes how risk assessment statutes create frameworks within which policymakers and technical actors are constrained and empowered when it comes to the design and implementation of a particular instrument. Specifically, it focuses on the tension between, on one hand, a top-down, global understanding of fairness, accuracy, and lack of bias, and, on the other, a tool that is well-tailored to local considerations. It explores three potential technical and associated policy consequences of SB 10's framework: proxies, Simpson's paradox, and thresholding. And it calls for greater attention to the design of risk assessment statutes and their allocation of global and local authority.

DOI: https://doi.org/10.15779/Z38WD3Q226

† 2017–19 PULSE Fellow, UCLA School of Law and 2019–20 Law Clerk, U.S. District Court for the District of Columbia. Alicia Solow-Niederman completed this work during her tenure as a PULSE Fellow, and the arguments advanced here are made in her personal capacity.

†† Ph.D Student, UCLA, Department of Computer Science.

††† Assistant Professor and Samueli Fellow, UCLA Department of Computer Science.

The authors are grateful to participants at the 2019 We Robot Conference and the March 2019 PULSE Workshop on AI and Law Enforcement. Thanks also to Solon Barocas, Chris Bavitz, Kristian Lum, Deirdre Mulligan, Richard Re, and the members of the UC Berkeley Algorithmic Fairness & Opacity Working Group for thoughtful comments and helpful conversations. This work is partially supported by NSF grants IIS-1657613, IIS-1633857, and CCF-1837129. Finally, many thanks to the editors of the *Berkeley Technology Law Journal*.

TABLE OF CONTENTS

I. INTRODUCTION

On August 28, 2018, California passed the California Money Bail Reform Act, also known as Senate Bill 10 (SB 10), and eliminated the state's system of money bail. Lauded by politicians as a "transformative" measure,[1] SB 10 aimed to deploy algorithmic pretrial risk assessment to combat socioeconomic disparities in the criminal justice system. But even groups that have long supported criminal justice reform criticized SB 10's final text, raising concerns about an "overly broad presumption of preventative detention" that compromised due process and racial justice.[2] Whether or not it makes sense to eliminate money bail is a policy decision beyond the scope of this Article. Still, any such policy decision must account for the fact that algorithmic pretrial assessments are not an objective substitute for subjective human considerations.

1. *See Governor Brown Signs Legislation to Revamp California's Bail System*, CA.GOV (Aug. 28, 2018), https://www.ca.gov/archive/gov39/2018/08/28/governor-brown-signs-legislation-to-revamp-californias-bail-system-protect-public-safety/index.html [https://perma.cc/772Z-YJ6C] (quoting California State Chief Justice Tani Cantil-Sakauye) ("This is a transformative day for our justice system. Our old system of money bail was outdated, unsafe, and unfair."); *see also id.* (quoting former Governor Jerry Brown) ("Today, California reforms its bail system so that rich and poor alike are treated fairly.").

2. *ACLU of California Changes Position to Oppose Bail Reform Legislation*, ACLU (Aug. 20, 2018), https://www.aclusocal.org/en/press-releases/aclu-california-changes-position-oppose-bail-reform-legislation [https://perma.cc/9MEP-9DXW].

Algorithmic risk assessment begins with the premise that an algorithm can provide a concrete measure of risk that informs a judge of salient facts about the defendant. But this premise is questionable, and a rapidly growing legal and technical literature has begun to underscore how and why data-driven algorithms are not automatically unbiased.[3] Building from a long-standing critique of actuarial assessments in criminal justice,[4] scholars, civil society members, and policymakers alike are contending with questions of bias and accountability,[5] competing definitions of fairness within such algorithms,[6] and concerns about the ways in which automated tools may reinforce underlying societal inequities.[7] Research to date tends to focus on fairness, accountability, and transparency[8] within the tools, urging technologists and policymakers to recognize the normative implications of these technical interventions.[9] While questions such as whether these

3. For instance, as an April 2019 report by the Partnership on AI emphasizes, "[a]lthough the use of these [algorithmic risk assessment] tools is in part motivated by the desire to mitigate existing human fallibility in the criminal justice system, it is a serious misunderstanding to view tools as objective or neutral simply because they are based on data." *Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System*, PARTNERSHIP ON AI 3 (Apr. 2019), https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/ [https://perma.cc/PN9D-68Q5].

4. *See* discussion *infra* notes 15–25.

5. *See, e.g.*, Vienna Thompkins, *What Are Risk Assessments—and How Do They Advance Criminal Justice Reform?*, BRENNAN CTR. (Aug. 23, 2018), https://www.brennancenter.org/blog/what-are-risk-assessments-and-how-do-they-advance-criminal-justice-reform [https://perma.cc/X8KS-CBCG]; Anna Maria Barry-Jester et al., *The New Science of Sentencing*, MARSHALL PROJECT (Aug. 4, 2015, 7:15 AM), https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing [https://perma.cc/99KU-Q4NC].

6. For an accessible overview of different technical definitions of fairness, see Arvind Narayanan, *Tutorial: 21 Fairness Definitions and Their Politics*, YOUTUBE (Mar. 1, 2018), https://www.youtube.com/watch?v=jIXIuYdnyyk [https://perma.cc/2VGP-M95T].

7. *See, e.g.*, Eric Holder, *Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference*, JUSTICE.GOV (Aug. 1, 2014), https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th [https://perma.cc/9E5C-MRSZ] ("By basing sentencing decisions on static factors and immutable characteristics . . . [risk assessment tools] may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.").

8. *See generally Conference on Fairness, Accountability, and Transparency (FAT*)*, FAT CONF., https://fatconference.org/index.html [https://perma.cc/8ASD-V9JG] (last visited Sept. 20, 2019).

9. By "normative," this Article broadly refers to the effect of an intervention on the common good and the life and liberty of individuals. *Cf.* Laurence Solum, *Legal Theory Lexicon: Welfare, Well-Being, and Happiness*, LEGAL THEORY BLOG (May 31, 2009), https://lsolum.typepad.com/legaltheory/normative_legal_theory/ [https://perma.cc/3CGY-F4PQ] ("[A]ny or most of the reasonable views about normative theory agree that what is good or bad for individual humans is morally salient.").

instruments are fair or biased are normatively essential, this Article contends that they cannot be the whole conversation. Automated risk assessment systems are not sterile tools that operate in a vacuum; rather, they are dynamic, norm-laden interventions that must be deployed within complex webs of new and preexisting policy requirements as well as legal institutions, rules, and associated social practices.[10] Accordingly, looking at issues such as fairness or bias in a tool in isolation elides vital bigger-picture considerations about the institutions and political systems within which tools are developed and deployed.[11]

Although SB 10 has been stayed pending a 2020 ballot referendum,[12] its statutory framework offers a real-world example of the political, legal, and

10. A growing number of nonprofit entities and academics recognize this point. *See, e.g.*, *Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System*, *supra* note 3, at 3 (articulating risk assessment challenges at three levels: "1. Concerns about the validity, accuracy, and bias in the tools themselves; 2. Issues with the interface between the tools and the humans who interact with them; and 3. Questions of governance, transparency, and accountability"); Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, 46 CRIM. JUST. & BEHAV. 185, 205 (2019) [hereinafter Eckhouse et al., *Layers of Bias*] (offering a framework to assess "layers of bias" in algorithmic risk assessment and emphasizing that every layer requires "human judgment about what fairness means"); Megan T. Stevenson, *Assessing Risk Assessment in Action*, 3 MINN. L. REV. 303, 306, 317 (2018) [hereinafter Stevenson, *Assessing Risk Assessment*] (underscoring "people and design choices" behind risk assessment and describing decisions about predicting risk, selecting a risk prediction algorithm, and dividing groups into classification levels as "choices that depend, at least partially, on the normative and legal landscape"); Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 61 (2017) (examining "normative judgments entailed in the development of predictive recidivism risk information").

11. Legal scholars have recently decried a lack of concrete evidence about risk assessment tools' efficacy and called for a more practical, empirically informed approach to evaluating risk assessment algorithms. *See* Stevenson, *Assessing Risk Assessment*, *supra* note 10, at 305–06; Brandon L. Garrett & John Monahan, *Judging Risk*, 108 CALIF. L. REV. (forthcoming 2019), https://ssrn.com/abstract=3190403 (manuscript at 30) [hereinafter Garrett & Monahan, *Judging Risk*] ("[F]ar more attention must be paid to the structure of decisionmaking that is supposed to be informed by risk assessment."). This Article concurs that attention to real-world outcomes, and not merely abstract risks, is imperative. It begins its human- and design-focused analysis one level up from the risk assessment tools themselves. This Article is the first account to evaluate the ways in which legislative choices about a risk assessment regime create a particular institutional context within which tools operate—and how this institutional context in turn structures the affordances and limitations of the tools.

12. SB 10 was temporarily stayed after a coalition of bail bond industry groups engaged the state's direct democracy system and placed a referendum on SB 10 on California's 2020 ballot. *See SB 10: Pretrial Release and Detention*, CAL. CTS., https://www.courts.ca.gov/pretrial.htm [https://perma.cc/3ZZF-GF5N] (last visited Sept. 20, 2019); *see also* Jazmine Ulloa, *Bail Bond Industry Moves to Block Sweeping California Law, Submitting Signatures for a 2020 Ballot Referendum*, LA TIMES (Nov. 20, 2018, 4:05 PM), https://www.latimes.com/politics/la-pol-ca-bail-referendum-signatures-20181120-story.html [https://perma.cc/QS93-8BZL].

statistical tradeoffs inherent in risk assessment statutes. This Article's detailed analysis of SB 10 focuses on a tension between, on the one hand, a top-down, *global* understanding of fairness, accuracy, and lack of bias, and, on the other hand, a tool that is well-tailored to *local* considerations. Anytime there is both a more centralized body that sets overarching guidelines about the tool and a risk assessment algorithm that must be tailored to reflect local jurisdictional conditions, as we will see is the case in SB 10, there will be a *global-local tension*. For instance, the pursuit of a single, statewide understanding of a first principle like non-discrimination—as consistency and rule of law might demand—requires technical tradeoffs in the fairness and accuracy of the tool. Risk assessment tools must be validated with reference to the particular conditions of application.[13] Yet such validation of a tool to make it more fair or accurate at a local level—as technological best practices demand—can produce different ground-level understandings of what fairness and accuracy require. And this local variability is in tension with a top-down, global understanding of the normative principle. In practice, any attempt to protect a normative principle within an algorithmic tool must be operationalized within the system of *algorithmic federalism* that we create each time a jurisdiction (typically a state) adopts risk assessment and then deploys the associated algorithmic risk assessment tool in its sub-jurisdictions (typically counties) that are responsible for discrete policy and technical steps in the instrument's creation and use.[14]

13. *See, e.g.*, PAMELA M. CASEY ET AL., OFFENDER RISK & NEEDS ASSESSMENT INSTRUMENTS: A PRIMER FOR COURTS 21 (2014) ("A local validation study will (a) inform any modifications that must be made to the content of the tool to optimize predictive validity in the local jurisdiction and ensure that it meets basic minimum scientific standards, and (b) inform the development of appropriate cutoff values for categorizing offenders into different risk levels."); Garrett & Monahan, *Judging Risk*, *supra* note 11 (manuscript at 40) ("Instruments should be re-validated over time, at reasonable intervals, and with attention to local variation in populations, resources, and crime patterns.") (internal citation omitted); John Logan Koepke & David G. Robinson, *Danger Ahead: Risk Assessment and The Future of Bail Reform*, 93 WASH. L. REV. 1725, 1757 (2018) [hereinafter Koepke & Robinson, *Danger Ahead*] ("For tools to make well-calibrated predictions from the start, they need to be trained on data that matches the conditions about which they are making predictions.").

14. Immense thanks to Kiel Brennan-Marquez for suggesting this term during an early presentation of this project. Akin to other federated decision-making bodies, "algorithmic federalism" refers to algorithmic systems that feature jurisdiction-wide authority over algorithmic policy but also include substantial space for regional or local bodies to accomplish whatever objective the algorithmic policy intervention seeks. For instance, SB 10 features both a statewide policy authority and county-wide bodies responsible for managing large portions of the technical implementation and contending with the policy outcomes. In future work (tentatively titled *Algorithmic Localism*), Alicia Solow-Niederman intends to apply existing scholarship on federalism and localism to explore local and global tensions, with an eye to the costs and benefits of allocating decision-making authority and discretion at different levels.

This Article uses SB 10 to explore a subset of these challenges, emphasizing particular technical issues that arise from the way that this statute allocates authority and discretion. It proceeds in four Parts. Part II first surveys the adoption of actuarial criminal justice tools in the twentieth century. It then canvasses recent state moves to implement algorithmic risk assessment tools as well as associated legal controversies and scholarly critiques. Part III describes SB 10 as an example of one state's plan to deploy risk assessment instruments. Specifically, it summarizes particular SB 10 provisions and their legislative history, focusing on how the provisions grant authority and discretion to institutional actors at both the state and local level. Part IV applies this analysis with a series of hypothetical narratives drawn from real-world demographic data in California. These narratives illustrate how even the best-intentioned actions can lead to unanticipated or undesirable results, given the way that a statute allocates authority to state and county-level actors. Part V considers how to design risk assessment statutes in light of the inevitable *global-local tension*. It closes by proposing that a policy choice to insert too many *layers of discretion* is likely to be problematic, no matter which tool is adopted, before offering several specific recommendations that could improve risk assessment statutes in general and SB 10 in particular.

II. RISK ASSESSMENT TOOLS

A. ACTUARIAL JUSTICE: A BRIEF HISTORY

Contemporary risk assessment instruments share a common heritage with far older criminal justice interventions. The link between old and new is the idea that public officers can make predictions about an individual's behavior that should inform the treatment of that individual today. This Section's stylized overview contextualizes contemporary algorithms as the latest iteration of this historic phenomenon.

For much of the twentieth century, choices about human liberty depended on obviously subjective factors. In the 1920s, parole boards began frequently invoking "crime prediction" in decisions about sentence length.[15] Factors included " 'the look in the prisoner's eye,' or [parole] board members' personal experiences, intuition, and biases."[16] And in making bail

15. *See* Thomas Mathiesen, *Selective Incapacitation Revisited*, 22 L. & HUM. BEHAV. 455, 458–59 (1998); Kevin R. Reitz, *"Risk Discretion" at Sentencing*, 30 FED. SENT'G REP. 68, 70 (2017) ("[P]rison sentence lengths in most U.S. jurisdictions are already based on predictions or guesses about offenders' future behavior, and this has been true—in multiple settings—for at least a century.").

16. Reitz, *supra* note 15, at 69 (citing DAVID J. ROTHMAN, CONSCIENCE AND CONVENIENCE: THE ASYLUM AND ITS ALTERNATIVES IN PROGRESSIVE AMERICA (1980);

and sentencing determinations, "clinical prediction," or "the largely unstructured clinical judgment of skilled practitioners," was used to assess the likelihood of recidivism.[17] Outside of the parole or pretrial context, moreover, police officers and agencies have long made choices about where to allocate limited resources based on risk assessment,[18] which is an inherently predictive enterprise.

Two significant types of changes occurred in the back half of the twentieth century. One significant shift was methodological. In the 1960s and 1970s, a growing sense that clinical predictions were unfairly subjective and hence susceptible to improper bias catalyzed evidence-based interventions.[19] The resulting tools were "actuarial."[20] Rather than relying on subjective expertise, these actuarial tools invoked statistics to "assign a quantitative risk score to an offender by assessing unalterable (e.g., static) individual factors (i.e., history of substance abuse and age at first offense) that have been statistically linked to the risk of recidivism in correctional populations and based on research involving large population samples."[21] In contrast to the paradigm shift from subjective to actuarial tools, subsequent developments have been more evolutionary. Over time, a third generation of statistical tools expanded beyond "static risk factors (such as criminal history, age, and gender)" to consider risks, needs, and "both static and dynamic risk factors such as educational status [and] employment."[22] As Kelly Hannah-Moffat explains, tools that rely on more dynamic factors are distinct because they

---

MARVIN E. FRANKEL, CRIMINAL SENTENCES: LAW WITHOUT ORDER (1973)); *see also* BERNARD HARCOURT, AGAINST PREDICTION, 7–18 (2007) (discussing parole boards' use of risk assessment instruments since the 1920s).

17. Kelly Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, 30 JUST. Q. 270, 274 (2013).

18. This Article adopts a common definition of risk assessment: "the process of using risk factors [factors that precede and statistically correlate with recidivism] to estimate the likelihood (i.e., probability) of an outcome occurring in a population." *See* Garrett & Monahan, *Judging Risk*, *supra* note 11 (manuscript at 7) (citing CARNEGIE COMM'N ON SCI., TECH., & GOV'T, RISK AND THE ENVIRONMENT: IMPROVING REGULATORY DECISION MAKING (1993)).

19. For discussion of studies reporting bias in human judgment, see Eckhouse et al., *Layers of Bias*, *supra* note 10, at 17–18.

20. As used here, "actuarial" refers broadly to empirically informed assessments, and thus contrasts with judgments based only on professionals' clinical decisions. *Cf.* NATHAN JAMES, CONG. RESEARCH SERV., R44087, RISK AND NEEDS ASSESSMENT IN THE FEDERAL PRISON SYSTEM 10 (2018) ("[I]t is argued that utilizing actuarial rather than clinical (i.e., professional judgment alone) risk assessment makes the process more objective and less susceptible to rater bias.").

21. Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, *supra* note 17, at 274. For a discussion of static tools' limitations, see *id.*

22. Garrett & Monahan, *Judging Risk*, *supra* note 11 (manuscript at 10); *see also* Hannah-Moffat, *supra* note 17, at 274–77.

"focus on treatment or rehabilitation of the offender to prevent reoffending, rather than simply predict recidivism . . . . This approach to risk differs importantly from the correctional use of static risk for preventive or selective incapacitation, diversion, or deterrence of recidivism through the administration of harsh penalties . . . ."[23] The fourth generation of tools continued to use various combinations of static and dynamic inputs while according more weight to the individualized needs of the defendant.[24] The fifth generation entails the application of machine learning techniques, discussed below, to provide more up-to-date predictions that take far more factors into account.[25]

Driven in large part by political forces and associated policy changes,[26] a second and partially overlapping shift attempted to limit what categories of inputs were permissible when imposing bail. The early 1960s witnessed mounting concern with overcrowded jails and the detention of less wealthy defendants, even when such individuals did not pose any public safety risk if they were released.[27] These reform efforts culminated with Congress's enactment of the 1966 Bail Reform Act, which aimed to "minimize reliance on money bail[,] . . . established that a defendant's financial status should not be a reason for denying their pretrial release, made clear that the risk of nonappearance at trial should be the only criterion considered when bail is assessed, and . . . generally forb[ade] judges from treating a defendant's dangerousness or risk to public safety as a reason for detention."[28]

But this initiative did not endure. In the 1970s and 1980s, the federal government undertook a fundamental reworking of the underlying reason for imposing bail.[29] Catalyzed by mounting public unrest in the civil rights era and the Nixon Administration's "tough on crime" stance, a second set of reforms effectively reversed the earlier policy. Rather than set terms that ensured a defendant's presence in court, pretrial detention emphasized the risk that the defendant would commit future crimes and threaten public safety.[30] In practice, as it emerged, this shift reframed the salient questions and asked judges to assess a defendant's perceived "dangerousness" to

23. Hannah-Moffat, *supra* note 17, at 276. For a more detailed account of risk/needs assessment in more modern tools, see *id.* at 274–76.

24. Garrett & Monahan, *Judging Risk*, *supra* note 11 (manuscript at 10).

25. *Id.*

26. This overview is indebted to John Logan Koepke & David G. Robinson's summary of this history. *See* Koepke & Robinson, *Danger Ahead*, *supra* note 13, at 1731–42.

27. *See* Garrett & Monahan, *Judging Risk*, *supra* note 11 (manuscript at 9).

28. *Id.* (citing United States v. Leathers, 412 F.2d 169, 171 (D.C. Cir. 1969)).

29. For a more detailed survey of changing bail practices in the United States, see Koepke & Robinson, Danger Ahead, *supra* note 13, at 1737–50.

30. Garrett & Monahan, *Judging Risk*, *supra* note 11 (manuscript at 10); *see also* Koepke & Robinson, *supra* note 13, at 1739–43.

determine the risk they posed.[31] Congress formally codified this evolution during the Reagan Administration with the 1984 Federal Bail Reform Act, which required federal judges to "consider danger to the community in all cases in setting conditions of release."[32] States largely followed the federal government's lead, marking a national shift in the discourse around risk assessment.[33]

The move towards assessing "dangerousness" goes hand-in-hand with the ongoing evolution of predictive risk assessment instruments. To date, risk assessment tools rely on relatively simple machine learning models, such as logistic regression, and do not yet embrace more complex machine learning models.[34] Machine learning operates by parsing large datasets to identify patterns in the data, which allows the development of a predictive function that can be applied to previously unseen data sets.[35] The use of an algorithm might be more standardized than the older clinical assessment based on, say, the look in the person's eye.[36] Yet it does not make the predictive enterprise objective; rather, it turns on a spate of human choices about what data to use, what statistical model to adopt, how to "tune" the model, and how to apply the findings.[37] As more complex machine learning methods are integrated into risk assessment instruments, it will become even more

31. Though beyond the scope of this paper, these developments occurred in tandem with a "selective incapacitation" movement that focused on detaining the most "dangerous" defendants. This effort dovetails with many of the objectives of recidivism-based risk assessment. For an analysis of contemporary lessons for algorithm risk assessment drawn from the history of selective incapacitation, see Danielle Kehl et al., *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*, RESPONSIVE COMMUNITIES INITIATIVE 3 (2017), http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041 [https://perma.cc/L7R9-K2W9].

32. United States v. Himler, 797 F.2d 156, 159 (3d Cir. 1986); *see also* Koepke & Robinson, *Danger Ahead*, *supra* note 13, at 1742.

33. *See* Koepke & Robinson, *Danger Ahead*, *supra* note 13, at 1740–41.

34. *See id.* at 1781 (suggesting that risk assessment is likely to progress from "logistic regression-based techniques toward more complex machine learning techniques"). Legal scholarship at times seems to distinguish between "real" machine learning and actuarial science based on logistic regression and other statistical methods. Rather than advance such a dichotomy, this Article positions contemporary risk assessment algorithms based on logistic regression as a simple machine learning model.

35. For discussion of machine learning, see generally VISHAL MAINI & SAMER SABRI, MACHINE LEARNING FOR HUMANS (2017). In supervised machine learning, the dominant contemporary method, the data scientist will "tune" different parameters to improve a selected statistical model's ability to deliver results that are closer to a predefined goal, or "objective function." *Id.*

36. *See* discussion *supra* text accompanying notes 15–18 and sources cited therein.

37. For an overview of the different steps necessary to arrive at a working machine learning model, see David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653 (2017).

essential to resist "automation bias" and ensure adequate oversight of the tool's fairness and accuracy.[38] As this Article underscores, the way that this predictive enterprise operates turns not only on individual choices, but also on initial policy choices about how to constrain or channel discretion and decisional authority.

B. ALGORITHMIC RISK ASSESSMENT: A RECENT HISTORY

The practical stakes are high because state and local jurisdictions in the United States are increasingly turning to algorithmic risk assessment. Though existing implementations span the pretrial and sentencing context, this Article focuses on pretrial risk assessment procedures like SB 10.[39] In the last seven years alone, half of U.S. states have either implemented or are seriously considering the use of some form of risk assessment tools in pretrial settings.[40] And many of these developments have been quite recent. According to the National Council of State Legislatures (NCSL),[41] in 2017

38. *Cf.* Danielle Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1271–72 (2008) ("The impulse to follow a computer's recommendation flows from human 'automation bias'—the 'use of automation as a heuristic replacement for vigilant information seeking and processing.' ") (quoting Linda J. Skitka et al., *Automation Bias and Errors: Are Crews Better Than Individuals?*, 10 INT'L J. AVIATION PSYCHOL. 85, 86 (2000)).

39. This narrower focus permits more detailed analysis of a particular set of interventions, without inadvertently conflating pretrial risk assessment and other forms of algorithmic criminal justice decisions, such as sentencing. However, this Article's broader conclusions about the importance of looking at the policy design, as well as its analysis of systems in which discretion is spread across global and local layers, are more generally applicable.

40. This figure is calculated from reports published by the National Conference of State Legislatures. *See Trends in Pretrial Release: State Legislation Update Civil and Criminal Justice* 1, NCSL (Apr. 2018), http://www.ncsl.org/portals/1/ImageLibrary/WebImages/Criminal%20Justice/pretrialEnactments_2017_v03.pdf [https://perma.cc/GYS8-HS55]; *Trends in Pretrial Release: State Legislation Update Civil and Criminal Justice* 1, NCSL (Mar. 2015), http://www.ncsl.org/portals/1/ImageLibrary/WebImages/Criminal%20Justice/NCSL%20pretrialTrends_v05.pdf [https://perma.cc/B88P-CHAK].

41. This research was funded in part by the Laura and John Arnold Foundation, now known as Arnold Ventures. This organization has supported the development of a widely used Public Safety Assessment (PSA) tool that has been adopted or is being implemented in over forty jurisdictions. *See Judicial Release Decisions*, ARNOLD VENTURES, https://www.arnoldventures.org/work/release-decision-making [https://perma.cc/G9CP-274Q] (last visited Sept. 20, 2019). Though beyond the scope of this Article, the lack of independent research or oversight of these tools by bodies that are not also invested in creating them is disconcerting. *Cf. The Use of Pretrial "Risk Assessment" Instruments: A Shared Statement of Civil Rights Concerns* 7, CIV. RTS. DOCS, http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-Full.pdf [https://perma.cc/86ZC-Q8DR] [hereinafter *A Shared Statement of Civil Rights Concerns*] (last visited Sept. 20, 2019) ("[A] pretrial risk assessment instrument must be transparent, independently validated, and open to challenge

alone, "[n]ine states enacted laws allowing or requiring courts to use risk assessments to assist in establishing bail and release conditions. Another five passed bills directing studies or development of risk assessment tools."[42] These enactments are, moreover, the latest in a longer-running series of state statutes that use these tools: "Since 2012, 20 laws in 14 states created or regulated the use of risk assessments during the pretrial process. In 2014 alone, 11 laws were passed to regulate how risk assessment tools are used to help determine whether, and under what conditions, a defendant should be released."[43]

Significantly, these enactments represent more than updates to technical instruments in order to keep up with the Joneses. There is also an underlying policy narrative. Like earlier historic shifts, these policies reflect a belief that more sophisticated algorithmic risk analysis tools can better account for individual defendant characteristics, rather than making a blanket choice to detain or release an individual based on their alleged charges.[44] And continuing debates that date to at least the 1960s, much of the discussion involves broader questions about the role of bail or non-monetary restrictions, as well as how these choices interact with fundamental rights and civil liberties. For instance, prior to California's elimination of money bail with SB 10, a 2017 New Jersey state statute changed New Jersey's money bail system to provide judges with algorithmic risk assessment scores, aiming to "build a better, fairer[,] and safer system of criminal justice."[45]

This apparent state enthusiasm for algorithmic solutions, however, has met mounting public and scholarly debate about the ethical and legal propriety of these tools. For instance, over a hundred civil society organizations recently signed and adopted a statement of civil rights concerns.[46] As this document details, concerns about the use of such tools

---

by an accused person's counsel. The design and structure of such tools must be transparent and accessible to the public.").

42. *Trends in Pretrial Release: State Legislation Update Civil and Criminal Justice*, *supra* note 40, at 1.

43. Amber Widgery, *Trends in Pretrial Release: State Legislation* 1, NCSL (Mar. 2015), http://www.ncsl.org/portals/1/ImageLibrary/WebImages/Criminal%20Justice/NCSL%20pretrialTrends_v05.pdf [https://perma.cc/7NM6-846B].

44. *See id.*

45. *See* Stuart Rabner, *Chief Justice: Bail Reform Puts N.J. at the Forefront of Fairness*, NJ.COM (Jan. 9, 2017, 9:33 AM), https://www.njcourts.gov/courts/assets/criminal/starledgercolumn.pdf [https://perma.cc/T9PK-LZV8].

46. *A Shared Statement of Civil Rights Concerns*, *supra* note 41; *see also More than 100 Civil Rights, Digital Justice, and Community-Based Organizations Raise Concerns About Pretrial Risk Assessment*, LEADERSHIP CONF. ON CIV. & HUM. RTS. (July 30, 2018), https://civilrights.org/2018/07/30/more-than-100-civil-rights-digital-justice-and-community-based-organizations-raise-concerns-about-pretrial-risk-assessment/ [https://perma.cc/M7V2-8X9P].

include, among others, the risk of data inputs that reproduce and reinforce racial inequities in the criminal justice system overall; the failure to provide adequate procedural safeguards, including individualized, adversarial hearings for all defendants; and the lack of transparency or access to the data or algorithms used by proprietary instruments.[47]

These long-simmering issues, moreover, began to boil over into more general consciousness with a 2016 ProPublica investigation alleging that a proprietary sentencing algorithm, COMPAS, was systematically unfair in its treatment of black defendants.[48] Specifically, ProPublica reported a problem with "error rate balance": COMPAS labelled more black defendants who did *not* reoffend as "high risk" (false positives) and more white defendants who *did* reoffend as low risk (false negatives).[49] This research was immediately met with rebuttals[50] contending that the tool was fair because it exhibited "predictive parity": at a given risk level, it predicted a roughly equal proportion of white and black reoffenders.[51] In addition to this technical fairness debate, which turns on how one defines fairness, others lodged critiques that no opaque, proprietary algorithm could be considered fair to a criminal defendant challenging it.[52] And still others, including the company that produced the tool, argued that any problem stemmed from the data

47. *See A Shared Statement of Civil Rights Concerns*, *supra* note 41. This Article reserves treatment of due process concerns and questions about whether these algorithms amount to unlawful preventative detention for separate work.

48. *See* Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/9857-ZXCZ]. For discussion of "the COMPAS debate," see Eckhouse et al., *Layers of Bias*, *supra* note 10, at 6–7.

49. Angwin et al., *Machine Bias*, *supra* note 48.

50. *See* Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions was Labeled Biased Against Blacks. It's Actually Not that Clear.*, WASH. POST (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ [https://perma.cc/ZSF4-QU6K]. *But see* Julia Angwin & Jeff Larson, *ProPublica Responds to Company's Critique of Machine Bias Story*, PROPUBLICA (July 29, 2016, 11:56 AM), https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story [https://perma.cc/5K54-2C2U].

51. *See* Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions was Labeled Biased Against Blacks. It's Actually Not that Clear*, *supra* note 50. For additional discussion of predictive parity, see *supra* text accompanying note 6.

52. *See, e.g.*, Matthias Spielkamp, *Inspecting Algorithms for Bias*, MIT TECH. REV. (June 12, 2017), https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/ [https://perma.cc/R6PD-LVCE] (discussing criminal defendant's argument that "due process was violated when the judge who sentenced him consulted COMPAS, because the workings of the system were opaque to the defendant").

itself and not the algorithm.[53] From this point of view, the salient factor is the unequal base rates of recidivism among racial groups in the observed population that makes up the machine learning dataset. Given such baseline differences, any tool that applies proportionate outcomes across the board at a particular risk level will have proportionately different effects on different racial groups. The debate about technical fairness thus connects to bedrock civil society concerns about racial inequity in the criminal justice system as a whole, even as proponents of these tools assert that they represent more objective, fairer ways to make criminal justice decisions.

Recent legal scholarship echoes these points. A growing literature cautions against algorithmic risk assessment as an automatic key to a fairer criminal justice system. In addition to critiques of the biased racial impact of risk assessment[54] and evaluations of what fairness means in practice,[55] scholars have begun to assess critically the different ways in which algorithmic systems may be unfair. For example, a recent article by Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini proposes three "layers of bias" that can arise in risk assessment models: first, whether the statistical model of fairness used is sound and desirable; second, whether the data itself contains embedded biases; and finally, whether it is fundamentally fair to make individualized criminal justice decisions based on statistical patterns derived from groups.[56] This third prong questions whether it is constitutionally valid to make criminal justice choices in this way, raising both equal protection and due process concerns.[57] Adding to the dialog around constitutional principles and new technical interventions, Brandon

53. *See* William Dieterich et al., *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, NORTHPOINTE (July 8, 2016), https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html [https://perma.cc/2TDA-25L8].

54. *See, e.g.*, Bernard E. Harcourt, *Risk as a Proxy for Race: The Dangers of Risk Assessment*, 27 FED. SENT'G REP. 237, 237 (2015) ("[R]isk today has collapsed into prior criminal history, and prior criminal history has become a proxy for race. The combination of these two trends means that using risk assessment tools is going to significantly exacerbate the unacceptable racial disparities in our criminal justice system.").

55. Again, there are myriad technical and ethical definitions that may be at odds with one another and with other values such as accuracy. For a video tutorial canvassing these issues, see Aaron Roth, *Tradeoffs Between Fairness and Accuracy in Machine Learning*, YOUTUBE (Jan. 30, 2017), https://www.youtube.com/watch?v=tBpd4Ix4BYM [https://perma.cc/EWD7-NL6D]; *see also* text accompanying notes 8–10 and sources cited therein.

56. *See* Eckhouse et al., *Layers of Bias*, *supra* note 10, at 1.

57. This branch of the literature relies on earlier work by scholars such as Sonja Starr, who has argued that actuarial risk assessments violate the Equal Protection Clause of the U.S. Constitution. *See* Sonja B. Starr, *Evidence-Based Sentencing and The Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014). *But cf.* sources cited *infra* note 105 (suggesting that contemporary Equal Protection jurisprudence is not an apt fit for algorithmic criminal justice).

Garrett and Jonathan Monahan have also raised constitutional questions involving judges' actual use of risk assessment algorithms, noting unsettled due process questions when a judge's determination is "informed by quantitative risk assessment methods."[58] As Garrett and Monahan acknowledge,[59] and as Megan Stevenson contends in another recent piece,[60] despite a bevy of theoretical concern about the use of these algorithmic tools, there is an extremely limited literature on their adoption in practice.

There is not only a lack of empirical evidence about the use of algorithmic tools on the ground, but also even less sustained attention to the design of statutes and regulations and the associated protocols, norms, and institutions within which risk assessment instruments are developed and deployed. This Article provides the first in-depth evaluation of ways in which choices about how to craft a policy and how to allocate discretion and authority within new and preexisting institutions inform both the theory and practice of algorithmic risk assessment.

III. SB 10'S STATUTORY STRUCTURE

This Part surveys SB 10 to illustrate how policymakers' choices about how to design risk assessment statutes in turn affect the creation and implementation of particular risk assessment instruments. It focuses on portions of SB 10 that grant authority and discretion to state and local actors, respectively.[61] This overview provides a foundation for Part IV's analysis of the specific policy and technical consequences of SB 10's statutory framework.

On its face, SB 10 grants local courts considerable discretion over the creation and implementation of the risk assessment instrument. The statute provides that local "Pretrial Assessment Services" (PAS) are responsible for pretrial risk level assessment of individuals who have been charged with a crime.[62] Each "particular superior court," which operates at the county level,[63]

58. Garrett & Monahan, *Judging Risk*, *supra* note 11, at 1–2.

59. *See id.* at 15–16.

60. *See* Stevenson, *Assessing Risk Assessment*, *supra* note 10, at 305–06.

61. This Article integrates proposed rules that California's Judicial Council has already published. Though it is possible that additional rules might have clarified ambiguities such as precisely what validation requires, they remain unresolved for the foreseeable future because further rule development is stayed pending the 2020 referendum. *See supra* note 12. Assumptions or inferences drawn from the available text are noted in line.

62. *See* S.B. 10, 2017–2018 Leg., Reg. Sess. §§ 1320.26(a), 1320.7(f), 1320.7(g) (Cal. 2018) (enacted) (codified at CAL. GOV'T CODE § 27771 and scattered sections of CAL. PENAL CODE).

63. There are 58 trial courts in California, one per county, known as superior courts. These courts have general jurisdiction over all state civil and criminal matters, unless

is to determine whether PAS in that county consists of "employees of the court, or employees of a public entity contracting with the court for those services."[64] Superior courts may opt into a regional consortium or multi-county PAS with an "adjoining county," with the limitation that "persons acting on behalf of the entity, division, or program shall be officers of the court."[65] Moreover, there is local control over who makes up the "court," which the statute defines to include " 'subordinate judicial officers,' if authorized by the particular superior court" in accordance with the California Constitution and Rules of Court.[66] Superior courts thus call the shots regarding creation of the PAS as a local institution.[67] This localized control continues after the creation of the PAS. Each PAS is to report risk assessment information to the trial court, along with recommended post-trial conditions of release.[68] These recommendations are non-binding, and each adjudicating judge retains discretion regarding the final pretrial decision.[69]

But this local discretion is not open-ended. Creating and using a risk assessment tool entails political and technical choices. And the decisions that must be made complicate the local control narrative. More specifically, within SB 10's statutory framework, global statutory provisions and statewide Rules of Court interact with and constrain the available set of local options. The relevant global actor for SB 10 is the California Judicial Council (Council), which acts as the "policymaking body" for the California court system. The

---

otherwise provided by statute or federal law. *See Superior Courts*, CAL. CTS., http://www.courts.ca.gov/superiorcourts.htm [https://perma.cc/M6LX-ASSD] (last visited Sept. 21, 2019). The number of judges per superior court varies by the size of the county. *Id.*

64. Cal. S.B. 10 § 4 (enacted) (adding CAL. PENAL CODE § 1320.7(g)).

65. *Id.*

66. *Id.* (adding CAL. PENAL CODE § 1320.7(a)).

67. Earlier versions of the statute called for a statewide oversight body to play a greater role in creation of the local PAS. *See* Sandy Uribe, Assembly Committee on Public Safety, *Senate Bill Policy Committee Analysis: SB 10 (Hertzberg)*, CAL. LEGIS. INFO. 5 (July 11, 2017), http://leginfo.legislature.ca.gov/faces/billAnalysisClient.xhtml?bill_id=201720180SB10 [https://perma.cc/P4GP-487S] (describing an "unnamed agency" that would be authorized "to oversee pretrial services agencies, to select a statewide pretrial assessment tool, to develop guidelines, and to provide training and assistance on pretrial release"). References to the "unnamed agency" and such top-down oversight were eliminated from the statute after the Chief Probation Officers of California expressed concern that SB 10 would "inhibit[] local control and flexibility relative to allowing each jurisdiction to determine who will handle the various parts of the pretrial program . . . at the local level." *Id.* at 15. The Judicial Council also expressed concern at an earlier stage of legislative development that SB 10 "would infringe on judicial discretion and independence." *Id.* at 13. Subsequent versions of the statute replaced the proposed "unnamed agency" with the current structure that combines the Council's oversight via rulemaking with increased local responsibility to ensure that the tools satisfy the standards set out in "scientific research."

68. Cal. S.B. 10 § 4 (enacted) (adding CAL. PENAL CODE § 1320.7(g))

69. *See, e.g.*, *id.* (adding CAL. PENAL CODE §1320.20(f)).

Council has long promulgated rules for the state's judicial system. Since its creation in 1926, it has operated under a state constitutional mandate to "improve the administration of justice."[70] In addition to general consideration for the public interest and judicial policymaking, the Council fulfills this mandate by setting official rules of court for the state.[71] The Council presently carries out its mission through a number of internal committees, advisory committees, and task forces, which generally include some combination of voting members and advisory members.[72]

In the SB 10 context, the Council manages the development of validated risk assessment tools, top-down. Specifically, the statute empowers the Council to "adopt California Rules of Court and forms" as needed to implement SB 10.[73] Among other responsibilities, the statute also directs the Council to:

- "Compile and maintain a list of validated pretrial risk assessment tools";[74]
- Identify, define, and collect "minimum required data to be reported by each court";[75]
- "Train judges on the use of pretrial risk assessment information when making pretrial release and detention decisions, and on the imposition of pretrial release conditions";[76] and

70. *See* CAL. CONST. art. VI, § 6; *see* also LARRY L. SYPES, COMMITTED TO JUSTICE: THE RISE OF JUDICIAL ADMINISTRATION IN CALIFORNIA 1 (2002), http://www.courts.ca.gov/documents/sipes_intro.pdf [https://perma.cc/4ZZM-RVS9]. The Council consists of twenty-one voting members who are assisted by advisory members and Council staff. It is meant to be responsive to the public as a whole, and not to any particular constituency.

71. *See* CAL. R. CT. 10.1(b).

72. For instance, one internal committee is the Rules and Projects Committee, which "establishes and maintains a rule-making process that is understandable and accessible to justice system partners and the public." CAL. R. CT. 10.10; *see generally Advisory Bodies*, CAL. CTS., http://www.courts.ca.gov/advisorybodies.htm [https://perma.cc/ZWH5-F9NV] (last visited Sept. 22, 2019). These advisory bodies are governed by the California Rules of Court. Cal. R. Ct. 10.30; *see also Judicial Council Governance Policies*, JUD. COUNCIL CAL. 4 (Nov. 2017), http://www.courts.ca.gov/documents/JCGovernancePolicies.pdf [https://perma.cc/W5UB-4LA3].

73. Cal. S.B. 10 § 4 (enacted) (adding CAL. PENAL CODE § 1320.24(a)).

74. *Id.* (adding CAL. PENAL CODE § 1320.24(e)(1)).

75. *Id.* (adding CAL. PENAL CODE §§ 1320.24(b)(1), 1320.24(e)(2)).

76. *Id.* (adding CAL. PENAL CODE § 1320.24(e)(3)).

- Consult with the Chief Probation Officers of California and "assist courts in developing contracts with local public entities regarding the provision of [PAS]."[77]

These provisions, read in a vacuum, may seem unproblematic. Some form of overarching oversight is intuitively appealing, particularly to the extent that a centralized authority like the Council can prevent local jurisdictions from relying on a tool that is biased, discriminatory, opaque, or otherwise problematic. Furthermore, the form of oversight prescribed in SB 10 may appear to simply emulate the relationship between the Council and local courts in other contexts.

In the algorithmic context, however, applying the same oversight strategy is insufficient at best and counterproductive at worst. First, centralized guidance from the Council does not guarantee uniform outcomes. Consider, for instance, how risk levels are determined. According to the statutory text, the Council is to appoint a "panel of experts and judicial officers . . . [that] shall designate "low," "medium," and "high" risk levels based upon the scores or levels provided by the instrument for use by [PAS]."[78] In other words, for each county, a PAS is to generate risk "scores or levels," and a statewide panel appointed by the Council is to designate which PAS "scores or levels" are associated with high, medium, or low risk levels.[79] The choice of risk threshold—which triggers the decision about how to treat an individual—is thus a global one.

Applying such a fixed global choice, however, may result in different outcomes in different local jurisdictions. As Part IV reveals, relying on the same low, medium, and high risk cutoffs in counties with different demographic distributions may produce different racial, gender, or socioeconomic effects, by county. The COMPAS debate about algorithmic unfairness arises in part from this point: if the baseline rate of arrest is

77. *Id.* (adding CAL. PENAL CODE § 1320.24(e)(4)).

78. *See id.* (adding CAL. PENAL CODE §1320.25(a), 1320.24(e)(7)) (directing the Judicial Council to "convene a panel of subject matter experts and judicial officers" to "designate 'low,' 'medium,' and 'high' risk levels based upon the scores or levels provided by the instrument for use by Pretrial Assessment Services"). This text is confusing as written because the PAS "risk score" may "include a numerical value or terms such as 'high,' 'medium,' or 'low' risk," yet the high/medium/low "risk level" is also to be set by the Council, based on the PAS risk score. *See id.* (adding CAL. PENAL CODE § 1320.25(a)).

79. *Id.* The statute does not clearly state whether the threshold set by the Council can be county-specific, or whether it must be uniform across the state. If it must be uniform across the state, then this would imply a coordinated calibration of the models, which would stymie local validation of the instrument. If it can be county-specific, then there would be more local variation regarding how the statute treats individuals at a given quantitative risk level. For further discussion, see *infra* Parts IV–V.

different for white and black defendants in two jurisdictions, and the same high/medium/low risk categorization applies globally, then different proportions of individuals from each race will be affected in each jurisdiction.[80] Reserving any normative critique of such a result, the practical upshot is the link between global and local authority. In a regime like SB 10 that relies on a centralized definition of risk levels, global policy choices set levels of risk at specified numerical points. These global decisions implement a particular technical understanding of fairness that controls local outcomes (as a policy matter), potentially without accounting for local differences (as a technical matter). This outcome is the byproduct of technical and policy constraints, as opposed to a consciously-pursued and explicitly stated definition of what is fair.[81]

Second, parsing these outcomes is all the more complicated in the case of SB 10 because other portions of the statute cut the other way: they grant more authority to localities, with less precise global guidance. Specifically, SB 10 specifies a presumptive outcome for individuals that PAS assesses as "high-risk" and "low-risk," subject to judicial override as well as a host of exceptions enumerated in the text.[82] For individuals deemed high and low

80. The COMPAS debate centers on the use of different base rates within a single county. As explored in detail *infra* Part IV, an important, yet underexplored, technical consideration involves the proper unit of analysis. This Article considers inter-county differences, including how different county-by-county racial demographics further problematize the initial decision of how to set global and local authority.

81. Some other risk assessment bills have been more explicit on this point. For example, Idaho House Bill No. 118 provided an explicit definition of fairness: " 'Free of bias' means that an algorithm has been formally tested and shown to predict successfully at the same rate for those in protected classes as those not in protected classes, and the rate of error is balanced as between protected classes and those not in protected classes." H.R. 118, 65th Sess., 1st Reg. Sess. (Idaho 2019). This draft thus required predictive parity among protected classes, although this language was removed from the final statute. *See* Act of July 1, 2019, ch. 258, 2019 Idaho Sess. Laws (2019). For a discussion of predictive parity, the measure adopted by the controversial COMPAS tool, see *supra* text accompanying note 6. The California statute does not contain such an explicit operationalization of what fairness requires. Whether the definition in the original Idaho draft is normatively desirable is beyond the scope of this Article.

82. *See* Cal. S.B. 10 § 4 (enacted) (adding CAL. PENAL CODE §§ 1320.20, 1320.10) (referring to high risk and low risk, respectively). As the Judicial Council's proposed rules explain:

> Prearraignment release of arrested persons will depend on their assessed risk level, determined by their score from the risk assessment tool and other information gathered from an investigation done by Pretrial Assessment Services, as follows:
>
> - Low risk: Pretrial Assessment Services must release persons assessed as low risk prior to arraignment, on their own recognizance except for those persons arrested for

risk, there are global rules.[83] Where PAS concludes that an individual is medium-risk, however, it is to recommend their release or detention according to "standards set forth in the local rule of court."[84] SB 10 provides in subsequent text that each superior court is to set these local rules of court "in consultation with [PAS] and other stakeholders."[85] The statute additionally authorizes local rules that "expand the list of exclusions for persons assessed as medium risk that [PAS] is not permitted to release," so long as some medium risk individuals are still released.[86] Apart from this requirement, the local courts are constrained only by the requirement that local rules are "consistent" with the Council's global rules of court.[87]

Though the relevant Council rules are in stasis for the foreseeable future,[88] the draft versions do not offer much more concrete guidance to superior courts. Draft Rule 4.40 provides that "[e]ach local rule must authorize release for as many arrested persons as possible, while reasonably assuring public safety and appearance in court as required."[89] Without more, however, these goals of "public safety" and "appearance in court" may not amply guide or constrain local actors. As Human Rights Watch warns in its

---

misdemeanors or felonies who fall within the exclusions listed in section 1320.10(e). (Pen. Code, § 1320.10(b).)

- Medium risk: Pretrial Assessment Services has authority to release on own recognizance or supervised own recognizance, or detain prearraignment, except for those persons subject to one of the exclusions listed in section 1320.10(e) or additional exclusions that may be included by a local court rule. (Pen. Code, § 1320.10(c).)
- High risk: Pretrial Assessment Services—and the court, if the court provides prearraignment review—is not authorized to release persons assessed as "high risk." Under sections 1320.10(e) and 1320.13(b), these persons must be held until arraignment when the court will make a release determination and set conditions of release, if applicable.

Criminal Procedure: Proper Use of Pretrial Risk Assessment Information; Review and Release Standards for Pretrial Assessment Services for Persons Assessed as Medium Risk, CAL. CTS. 2 (2018), https://www.courts.ca.gov/documents/SP18-23.pdf [https://perma.cc/77EA-G2UR] (internal citations omitted) [hereinafter *Proposed Rules 4.10 & 4.40*].

83. *See id.*

84. Cal. S.B. 10 § 4 (enacted) (adding CAL. PENAL CODE § 1320.10).

85. *See id.* (adding CAL. PENAL CODE § 1320.11(a)).

86. *See id.*; *see also Proposed Rules 4.10 & 4.40*, *supra* note 82, at 13 ("If a court chooses to add to the list of exclusionary offenses or factors, the court must not adopt a rule that includes exclusions that effectively exclude all or nearly all persons assessed as medium risk from prearraignment release.").

87. Cal. S.B. 10 § 4 (enacted) (adding CAL. PENAL CODE § 1320.11(a)).

88. *See supra* note 12.

89. *Proposed Rules 4.10 & 4.40*, *supra* note 82, at 13.

comments on this proposed rule, "[t]his statement of purpose needs some specific regulations to make it meaningful."[90] Might there be reasons for local tailoring of these policy choices? Perhaps. Yet there are also institutional tradeoffs when global rules rely on local determinations to craft the policy. A framework that requires too much localization to craft the basic rules of the policy will not be globally consistent, and that lack of uniformity might itself be seen as unfair.

Regardless of the equilibrium that is ultimately struck, when human life and liberty are at stake, these choices should be made intentionally, with an awareness of the tradeoffs. Without designing a statute to account for these global-local tensions and tradeoffs, we risk encoding these understandings implicitly, in ways that are opaque and may resist democratic accountability. But the ways that risk assessment algorithms interact with systems of local and global discretion presently appear to be the inadvertent consequences of particular policy choices—not intentionally selected outcomes. With an emphasis on global-local tensions, Part IV surveys some of the ways in which implementing SB 10 as a technical matter entails policy judgments that are not explicitly specified by the statute.

IV. ALGORITHMIC RISK ASSESSMENT IN PRACTICE: POTENTIAL CONSEQUENCES OF THE SB 10 FRAMEWORK

Using SB 10 as an example of a risk assessment statute, this Part considers how a statistically-driven risk assessment tool might operate in practice. This Part specifically focuses on how SB 10's framework might create three categories of technical issues: proxies, Simpson's paradox, and thresholding. This detailed analysis and associated modelling not only highlights potential specific risks of SB 10, but also illustrates more generally the manner in which a statute's initial allocation of global and local authority will produce unexpected technical—and associated policy—consequences.

The first category, proxies, refers to features that would typically be considered valid for risk assessment, yet which are correlated with protected attributes that may not be considered for legal or policy reasons. SB 10 does

90. *Human Rights Watch Comments on California Judicial Council Bail Reform Rules*, HUM. RTS. WATCH (Dec. 10, 2018, 9:00 AM), https://www.hrw.org/news/2018/12/10/human-rights-watch-comments-california-judicial-council-bail-reform-rules [https://perma.cc/79U9-KSAA]; *see also Written Comments of the Electronic Frontier Foundation on Proposed Rules 4.10 and 4.40—Invitation to Comment* #SP18-23, ELECTRONIC FRONTIER FOUND. 15 (Dec. 14, 2018), https://www.eff.org/files/2018/12/18/written_comments_of_the_electronic_frontier_foundation_on_proposed_rules_4.10_and_4.40_invitation_to_comment_sp18-23_december_14_2018.pdf [https://perma.cc/7HSC-T8TB].

not contain an explicit discussion on how to handle the proxy problem. As we will see, this issue is likely to arise in risk assessment statutes more generally because neither local discretion nor a more centralized policy provides a complete solution.

The second category, Simpson's paradox, is a phenomenon in which the existence or the direction of a statistical trend differs between a global population (e.g., state) and its local sub-populations (e.g., counties). This issue is particularly acute in a statutory framework like SB 10, which relies on allocation of authority across both the global and local levels.

The final category, thresholding, refers to the process of setting qualitative risk levels associated with particular quantitative risk scores. SB 10 outlines a process in which the scores from a risk assessment tool are thresholded to group defendants into low, medium, and high risk levels, with associated policy outcomes. The policy rub is that thresholding can lead to unfair categorizations even when using scores that the relevant authority has deemed unbiased, especially where different protected groups are geographically concentrated or dispersed in different areas.

As the remainder of this Part reveals, a statutory framework like SB 10 matters because it will affect how each of these technical challenges arise and shape the available solution set.

A. PROXIES

As a policy matter, SB 10 and the publicly available draft rules address fairness both globally and locally.[91] Globally, the California Board of State and Community Corrections is to contract with independent third parties to assess the statute's effects with an emphasis on "the impact of the act by race, ethnicity, gender, and income level."[92] Locally, the implementing court is to "consider any limitations of risk assessment tools in general, and any limitations of the particular risk assessment tool used by [PAS], including . . . [w]hether any scientific research has raised questions that the particular instrument unfairly classifies offenders based on race, ethnicity, gender, or income level."[93]

Since the global requirements refer to independent audits, this Section emphasizes local fairness oversight: What might local consideration of risk assessment limitations require, and what research might be relevant to determine whether there has been unfair classification? The answer turns on technical design decisions about what kinds of information a tool can take into account. Suppose that a risk assessment instrument eliminates any use of

91. This analysis reflects rules made publicly available as of late May 2019.
92. Cal. S.B. 10 § 4 (enacted) (adding CAL. PENAL CODE § 1320.30(a)).
93. *Proposed Rules 4.10 & 4.40*, *supra* note 82, at 11–12.

a sensitive attribute, such as race, in making predictions about whether a given individual will recidivate. As the growing fairness literature documents, this solution does not guarantee that the tool is free of bias based on that attribute.

The problem is one of *proxies*,[94] a phenomenon wherein other valid features can be highly correlated with the protected attribute. Where a proxy for a protected attribute exists, decisions that incorporate a proxy variable may be biased, even when the decision does not explicitly incorporate the protected attribute.[95] Consider a non-technical example: the problematic historic practice of "redlining" a neighborhood, thereby permitting racial discrimination without explicit consideration of race.[96] Or at the individual level, a person's name can be a proxy for gender, such that making a decision about Aaron versus Erin could permit gender discrimination without explicit consideration of gender.

Unfortunately, simply forbidding the use of proxies for protected attributes is not a viable solution. Consider a feature like education. Information about an individual's education, such as highest degree obtained and undergraduate major, is correlated with gender,[97] such that education acts

94. Here, a proxy variable is a feature that can be used to infer one or more of the protected attribute values of an individual. A well-known example proxy for race is zip code.

95. *See* Eckhouse et al., *Layers of Bias*, *supra* note 10, at 15–16. Eckhouse stated:

> [E]ven a determined effort to exclude proxies for race, class, or other marginalized categories [from an algorithm] is not likely to be successful . . . . [O]mitting race from the set of variables in the original data set does not mean race is not included in the analysis; it merely induces remaining variables that are correlated with both race and the outcome variable to behave as if they are, in part, proxies for race.

Id. (citing Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 721 (2016); Cynthia Dwork et al., *Fairness Through Awareness*, PROCS. 3RD INNOVATIONS THEORETICAL COMPUTER SCI. CONF., 214 (2012); Devin G. Pope & Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 AM. ECON. J.: ECON. POL., 206–31 (2011)).

96. "Redlining" refers to the 1930s federal Home Owners' Loan Corporation's practice of marking neighborhoods in green, blue, yellow, or red to demarcate their credit risk. Under this schema, "[t]he 'redlined' areas were the ones local lenders discounted as credit risks, in large part because of the residents' racial and ethnic demographics. They also took into account local amenities and home prices." Tracy Jan, *Redlining Was Banned 50 Years Ago. It's Still Hurting Minorities Today*, WASH. POST (Mar. 28, 2019), https://www.washingtonpost.com/news/wonk/wp/2018/03/28/redlining-was-banned-50-years-ago-its-still-hurting-minorities-today [https://perma.cc/HU68-WAWK]. This practice allowed loans to be systematically denied to residents of minority-dominated neighborhoods by invoking credit risk without explicitly referring to race or ethnicity. Such discrimination was possible because, as operationalized, credit risk was in fact a proxy for race.

97. For a visualization of this data, which is provided by the U.S. Census Bureau, see *Detailed Educational Attainment Sex Ratio*, STAT. ATLAS (Sept. 4, 2018),

as a proxy for gender.[98] However, it is not feasible to dismiss it entirely as a factor in decisionmaking because it may in fact be relevant to the decision. For instance, the highest degree obtained or an individual's specialty area could be highly salient in hiring an individual. A proxy variable, in short, may include crucial information for making predictions.[99] Accordingly, rather than omitting a proxy variable entirely, it may be more valuable to exploit it for prediction while still limiting the bias due to its correlation with the protected attributes. In fact, several works have suggested that the only way to make a system truly fair, even from the effects of proxies, is to collect and take into account the protected attribute values at the learning or prediction stage.[100]

SB 10's framework allocates global and local authority in a way that can produce several proxy-related issues. Recall that the Council is to "[c]ompile and maintain a list of validated pretrial risk assessment tools,"[101] yet a county-level superior court and the associated PAS are to ensure that the tool used in their particular jurisdiction does not unfairly classify. At the same time, assuming that the Council's rules regarding validation are in line with technical best practices, any such rules must provide for local validation and

---

https://statisticalatlas.com/state/California/Educational-Attainment#figure/detailed-educational-attainment-sex-ratio [https://perma.cc/4NJR-LEQ2].

98. Informally, correlation or dependence indicates whether a proxy variable can be used to predict one or more protected attributes. Approaches to quantify this relationship include the Pearson correlation coefficient, which measures a linear relationship between variables, and mutual information, which measures probabilistic dependence between variables. Although not explored in detail here, proxies also raise practical difficulties. Determining both the correct statistical definition to quantify the proxy effect and whether a variable is an unacceptable proxy, whether accomplished through numerical thresholds or qualitative assessment, demands substantial time and resources. These costly investments decrease the efficiency gains of a turn to algorithmic risk assessment.

99. For an excellent discussion of proxies in the disparate impact context, see Barocas & Selbst, *supra* note 95, at 720–22.

100. Some approaches include learning new representations of the non-protected attributes such that they still have predictive power while remaining independent of the protected attribute. These strategies generally require accounting for the protected attribute values at both the learning and prediction stage. At the learning stage, the restriction is placed on the data used to develop the tool. At the prediction stage, the restriction is on the features that the tool gets as input to make a prediction. *See* Richard Zemel et al., *Learning Fair Representations*, PROC. 30TH INT'L CONF. ON MACHINE LEARNING 325 (2013); Michael Feldman et al., *Certifying and Removing Disparate Impact*, PROC. 21st ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 259 (2015). Moreover, it is possible to make a learned tool fair by setting protected attribute specific threshold rules. *See* Sam Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, PROC. 21st ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 797 (2017).

101. S.B. 10, 2017–2018 Leg., Reg. Sess. § 1320.24(e)(1) (Cal. 2018) (enacted) (codified at CAL. GOV'T CODE § 27771 and scattered Sections of CAL. PENAL CODE).

adaptation of a tool that is on the approved list.[102] Under this structure, each local superior court remains responsible for validating the risk assessment instrument used in its respective county.

To see how this allocation of authority might operate, consider applying SB 10's structure to a hypothetical state of Idem, made up of Counties A, B, and C. The counties are identically distributed for all observable measures.[103] They begin with the same risk assessment tool, selected from the list that Idem's statewide council has provided. Assume further that the counties are able to tweak the preapproved tool during the validation process.[104] They each proceed to validate and adapt the instrument, with an eye to ensuring it does not unfairly classify on the basis of the protected characteristic of race, as the statute requires.

Suppose County A and County B, when acting in good faith to comply with the state's requirement to avoid unfair classification when applying this tool, are each concerned that the use of any racial information at all is problematic.[105] They therefore decide to forbid risk assessment algorithms

---

102. These validation guidelines were not completed before the statute was stayed. *See* sources cited *supra* note 13 and accompanying text.

103. Here, observable refers both to demographic characteristics and available information, such as past criminal history.

104. Again, this point is ambiguous on the face of SB 10 without more explicit rules regarding validation, which are not yet published. This assumption nonetheless illustrates how the very choice of where and in what ways to permit validation, globally or locally, is a policy decision with technical implications that tend to be underexplored at the policy design stage.

105. County A's concern could apply even if the officials are not acting with discriminatory purpose such that the U.S. Constitution's Equal Protection Clause would bar the consideration of race. As Sam Corbett-Davies and Sharad Goel explain:

> [T]he dominant legal doctrine of discrimination focuses on a decision maker's motivations. Specifically, equal protection law—as established by the U.S. Constitution's Fourteenth Amendment—prohibits government agents from acting with 'discriminatory purpose' (Washington v. Davis, 1976). It bars policies undertaken with animus (i.e., it bars a form of taste-based discrimination, since acting with animus typically means sacrificing utility); but it allows for the limited use of protected attributes to further a compelling government interest (i.e., it allows a form of statistical discrimination).

Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, ARXIV 4 (Aug. 14, 2018), https://arxiv.org/abs/1808.00023 [https://perma.cc/CG8B-WNZJ]; *cf.* Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1053 (2019) (articulating limitations of "equal protection jurisprudence in relation to algorithmic criminal justice" and offering that this jurisprudence "is not a coherent or morally acute metric"). For purposes of the above hypothetical, assume that there is no applicable federal statutory regime that supports a disparate impact analysis, that the state constitution does not outright bar discrimination or preferential treatment on the

from using race as a feature and also bar the consideration of proxies that might correlate with race, such as zip codes. However, they make different choices about what constitutes an unacceptable proxy. Each county makes these decisions based on the following hypothetical distribution in which white individuals are more likely to have a high school diploma than black individuals (sixty percent vs. forty percent), such that education level is correlated with race:

Table 1. Proportionate size and high school completion rate per racial group.

	Proportion of Population	H.S. Diploma
White	0.5	0.6
Black	0.5	0.4

Based on this finding, County A bars the use of education level in the tool. On the other hand, County B does not require that the tool be completely independent of an individual's education level, on the grounds that it provides salient information about the risk that an individual poses. Lastly, County C permits the use of race as a feature—but the local PAS that administers the tool stipulates that it can only be used to correct for possible inadvertent bias from proxies. In other words, County C explicitly attempts to use race to achieve demographic parity, whereas the other counties bar the use of race or proxies for race. What happens when each county's tool is applied to a demographically identical individual?

The result varies by county. First, take County A. Recall that County A does not permit the use of race or education level as a feature. The tool it uses thus outputs scores such that 0.5 of the defendants are detained, at the average risk rate of the entire population, regardless of their race or education level,[106] as summarized in the following table:

---

basis of a suspect classification like race or national origin, and that there is no explicit racial animus that could establish a federal constitutional violation under *Washington v. Davis* and its progeny. A state actor might nonetheless decide to avoid the use of a particular characteristic like race or gender. At least one recently-enacted statute presently takes a hardline approach of this sort and bans consideration of gender. *See* Ann Carrns, *In California, Gender Can No Longer Be Considered in Setting Car Insurance Rates*, NY TIMES (Jan. 18. 2019), https://www.nytimes.com/2019/01/18/your-money/car-insurance-gender-california.html [https://perma.cc/9AKK-U9PD] (quoting California Insurance Department: " 'Gender's relationship to risk of loss no longer appears to be substantial,' the department noted, saying the rationale for using it was 'suspect' ").

106. For the sake of simplicity and clarity, we assume here that the risk assessment tool itself makes a decision whether to detain or release a defendant. In practice, the tool outputs

Table 2. Proportion of individuals in State of Idem who pose risk, by racial group and education level.

	No H.S. Diploma	H.S. Diploma	Combined
White	0.825	0.25	0.48
Black	0.62	0.375	0.52
Combined	0.7	0.3	0.5

Although it can be considered fair because it makes no distinction between individuals by their race or education level, such a tool sacrifices accuracy as an assessment algorithm. Next, consider County B, which bars the explicit use of race but permits some consideration of education level. Such a tool can infer an individual's race through their education level. For example, as Table 2 illustrates, the average risk rate for black defendants with no high school diploma is 0.62, but a failure to account for race and to use the combined risk score would mean that black individuals with no high school degree will be detained at a higher rate of 0.7. Finally, take County C, which uses race to correct for inadvertent bias from a proxy (here, education level). By considering both an individual's race and education level and thereby explicitly accounting for proxies, County C will achieve demographic parity in the manner illustrated below in Table 3.

Table 3. Rate of detention determined by tool in County C, by racial group and education level.

	No H.S.	H.S. Diploma	Combined
White	0.85	0.27	0.5
Black	0.6	0.35	0.5

In County C, the same black individual with no high school diploma will face a 0.6 likelihood of being detained, a proportion that is lower than that in County B. Combining these three examples, the result is that three counties using the same instrument can each attempt to account for "any limitations of the particular risk assessment tool used by [PAS], including . . . [w]hether any scientific research has raised questions that the particular instrument unfairly classifies offenders based on race, ethnicity, gender, or income

risk scores, which are then categorized using thresholds set by the global Council. *See* discussion *infra* Section IV.C.

level."[107] Yet each will treat an identical individual differently because they interpret unfair classification differently in their treatment of proxies.

This outcome, moreover, illustrates a broader global-local proxy tension. As a technical matter, a policy that permits each locality to determine how to handle proxies permits inconsistent decisions regarding treatment of proxies across counties. It is thus possible to have different outcomes for the same hypothetical individual based on the county where they are located, which may feel unfair from the perspective of the individual. But as a policy matter, county-level discretion about how to validate a tool may enhance local self-determination, which may feel more fair from the perspective of the locality as a whole. And the issue is even more complex because there may also be important technical reasons to permit local proxy determinations.

To see why, imagine that the global statute outright barred any use of particular protected attributes or their proxies. The trouble with such a global fiat is that the strength of the proxy, defined as its level of correlation with the protected attribute, can vary both in each county and in the state as a whole. In such a case, a county that does *not* exhibit a correlation between a protected attribute and a specified proxy would need to exclude variables that do not pose a huge proxy problem, thereby sacrificing accuracy. On the other hand, county-by-county determination of proxies is likely to be not only extremely resource intensive, but also problematic insofar as the result is a patchwork of county-by-county policy calls that make global oversight challenging. Neither a global nor a local solution is perfect. And as we will see in the following two Sections, global policy requirements come with further costs.

B. SIMPSON'S PARADOX

Without a uniform, top-down understanding of what fairness requires, problems might arise where global as opposed to local parsing of data leads in different directions. Specifically, applying SB 10's requirements, there is a risk of *Simpson's paradox*. Simpson's paradox refers to a statistical phenomenon in which a trend appears in several subgroups of a population, yet that same trend disappears or reverses when the subgroups are aggregated.[108] Consider, for instance, a group of men and women who work at a university that has two departments: Department A and Department B. Imagine a hypothetical algorithm that determines who receives a positive outcome, such as promotion in the department. In Department A, twenty-five percent of women receive a favorable decision, and no men receive that decision, such that there is evidence of systematic bias against these men. In

107. *Proposed Rules 4.10 & 4.40*, *supra* note 82, at 11–12.

108. Judea Pearl, *Understanding Simpson's Paradox*, 68 AM. STATISTICIAN 8 (2014).

Department B, 100% of women receive a favorable decision, and seventy-five percent of men do. Again, there is evidence that men are systematically less likely to receive the desirable result, even though the outcome is less skewed. Across the university as a whole, however, Table 4 outlines how there is complete gender parity, such that global oversight alone would not reveal any evidence of unfairness.

Table 4. Proportion of individuals who receive positive decision in each subgroup and respective population size.

	Proportion		Population Size	
	Women	**Men**	**Women**	**Men**
Dept. A	0.25	0.00	40	20
Dept. B	1.00	0.75	20	40
Combined	0.50	0.50	60	60

This effect can also be easily found in real-life data. For instance, consider the following real-world data on education level and ethnicity from seven California counties.[109]

Table 5. Proportion of individuals with four-year college degree or higher and population size (in thousands) by county.

	Proportion		Population Size	
	White	**Asian**	**White**	**Asian**
Fresno	**0.30**	0.28	219.3	53.3
Lassen	**0.16**	0.07	15.9	0.3
Marin	**0.62**	0.59	144.3	11.4
Monterey	**0.42**	0.36	104.6	16.5
San Mateo	**0.56**	0.55	236.3	147.8
Sutter	**0.22**	0.21	33.0	9.2
Tuolumne	**0.22**	0.11	34.2	0.3
Combined	0.44	**0.47**	787.6	238.8

109. This data covers 2011–2015 and relies on figures made available by the California Department of Public Health. *See Educational Attainment*, CHHS OPEN DATA, https://data.chhs.ca.gov/dataset/educational-attainment [https://perma.cc/YCL8-PHG7] (last visited Sept. 23, 2019).

As Table 5 shows, the attainment rate of a four-year college degree in each of seven California counties is higher among the white population than among the Asian population, but the trend is reversed when these counties are grouped together.

Turning back to algorithmic fairness, SB 10 would allow Simpson's paradox to occur if there is an opportunity for discrimination against a group at the county level, yet the discrimination becomes undetectable or even reverses direction at the state level, or vice versa.

For example, an individual's education level is one factor that contemporary instruments often use.[110] How might this consideration play out in the SB 10 context? Suppose that at least some of the tools approved by the Council take education level into account. Suppose, further, that the seven counties shown above decide to adopt risk assessment instrument E, which considers education as a major feature. For the sake of argument, say that tool E assigns a low numerical risk score to individuals with a four-year college degree or higher, and that this assessment holds across each of the seven counties. Assume that this quantitative risk score falls into the "low risk" qualitative category for the state. In other words, across the state, tool E's assessment leads to the conclusion that individuals who have completed four or more years of higher education are, collectively, a low risk group.

But this statewide result will not necessarily lead to uniform outcomes across each of the counties in the state; to the contrary, such a scenario could easily produce differential risk assessments for members of different demographic groups. Imagine, first, that an independent statewide auditor wishes to confirm that the tool is not unfairly classifying based on ethnicity. Given the cost and difficulty of validating with reference to each individual county, such an entity is likely to validate the tool using the aggregate data alone. Based on aggregate data, the tool could be rejected for discriminating against the white population on the ground that the 0.03 difference between the respective white and Asian assignment rates to the low-risk group is

110. For example, COMPAS, the tool that was used in the *Loomis v. Wisconsin* case, considers education level. *See* PAMELA M. CASEY ET AL., OFFENDER RISK & NEEDS ASSESSMENT INSTRUMENTS: A PRIMER FOR COURTS, NAT'L CTR. STATE CTS. A-21 (2014). The Federal Pretrial Risk Assessment (PTRA) uses highest educational attainment as one of the features of its logistic regression model. *See* Christopher T. Lowenkamp & Jay Whetzel, *The Development of an Actuarial Risk Assessment Instrument for U.S. Pretrial Services*, 73 FED. PROBATION 2 (2009), https://www.uscourts.gov/sites/default/files/73_2_3_0.pdf [https://perma.cc/D23V-HYRY]. Several counties in Maryland also rely on locally-developed tools that take education level into account. *See* Angela Roberts & Nora Eckert, *As Maryland Courts Meld Artificial Intelligence into Bail Decisions, Concerns Follow*, CAP. NEWS SERV. (Dec. 21, 2018), https://cnsmaryland.org/interactives/spring-2018/plea-bargain/pretrial-riskscore.html [https://perma.cc/7QA2-K6ZB]; *cf.* Eaglin, *supra* note 10, at 81 (discussing factors, including education level, used in existing tools).

unacceptable. Alternatively, it could be accepted on the grounds that the 0.03 difference between the white and Asian population in being classified as low risk is negligible. Each of these conclusions is a normatively-laden policy determination, embedded in the auditing process, that merits explicit, ex ante consideration by scholars and policymakers.

Critically, even in a world where there exists global consensus on what represents an acceptable difference across demographic groups, Simpson's paradox remains at the county level. In the case of the seven California counties, for instance, a tool deemed globally fair would in fact be biased against Asians in every one of the counties. To see how, return to the hypothetical tool E described above, for which achieving a four-year degree mediates the quantitative risk assessment level and results in a low risk label. Applying such a tool to counties with the above demographics, Asians would receive comparatively higher risk scores than whites within each of the counties. Furthermore, such discrimination can be much more significant than the 0.03 difference observed at the state level. As the above chart illustrates, in Tuolumne for instance, white individuals are twice as likely to be in the low-risk group than Asian individuals. Accordingly, different local and global parsing of the same data set can make even the same outcomes seem fair or unfair, depending on which vantage point is adopted.

The existence of Simpson's paradox thus provides a hefty challenge to SB 10's validation process. While the counties are responsible for developing fair risk assessment tools at the county level, the state must also validate each tool. If the state chooses to validate tools using statewide aggregated data, then Simpson's paradox may mean that the evaluation of a tool's fairness at the state level differs from the evaluation of that same tool at the county level. On the other hand, the state's validation of a tool for each individual county is not only expensive, but also an oxymoron if the idea is a centralized, uniform validation process set forth at the state level. Moreover, Simpson's paradox can interact with, and further complicate, the proxy issues described previously. In particular, determining whether a feature is a proxy for a protected attribute involves measuring the statistical correlation, which may hide itself or even reverse direction at the state level as opposed to county level, and vice versa. There is no panacea for these global-local challenges. The only palliative is more awareness and intentionality in making initial choices about which authority is responsible for which steps, along with greater upfront specification about who is to make determinations at which points in the process.

C. THRESHOLDING

Specific policy provisions within SB 10 also carry unrecognized technical—and human—consequences. Recall that the statute requires the

global Council to designate which PAS "scores or levels" are associated with high, medium, or low risk levels.[111] In operational terms, this means that each individual is first assigned a score by a risk assessment tool. Then, the individual is categorized into a risk group by thresholding their score: high risk if the score is above a certain value, low risk if it is below another value, and so on. It may be tempting to assume that for any scores generated by a fair system, however fairness is defined, a given threshold will guarantee fairness.

But the practice of thresholding itself can produce a new set of challenges.[112] First, as the technical fairness literature has emphasized, there are several mathematical notions of fairness, each of which may require a different set of thresholds.[113] For example, many notions of fairness aim to achieve a balance of classifier performance metrics, such as positive predictive value, false negative rate, and false positive rate,[114] between different protected groups. Scholars have shown that no single threshold rule can satisfy all three of these fairness definitions except in an unlikely assumption;[115] thus, choosing a threshold to enforce one notion of fairness would violate another.

This Article emphasizes a less recognized second point: even with an agreed-upon definition of fairness, enforcing it among subpopulations, such as different ethnic groups, may be impossible without setting a different threshold for each group.

---

111. Despite some textual ambiguity, the choice of the risk threshold—which triggers the decision about how to treat an individual—is up to the Council. *See supra* note 79. Again, the present discussion assumes that SB 10 provides for levels that are uniform across the state.

112. For an interactive visualization of thresholding challenges, see Martin Wattenberg et al., *Attacking Discrimination with Smarter Machine Learning*, GOOGLE RES., http://research.google.com/bigpicture/attacking-discrimination-in-ml/ [https://perma.cc/A6JQ-496Q] (last visited Sept. 23, 2019).

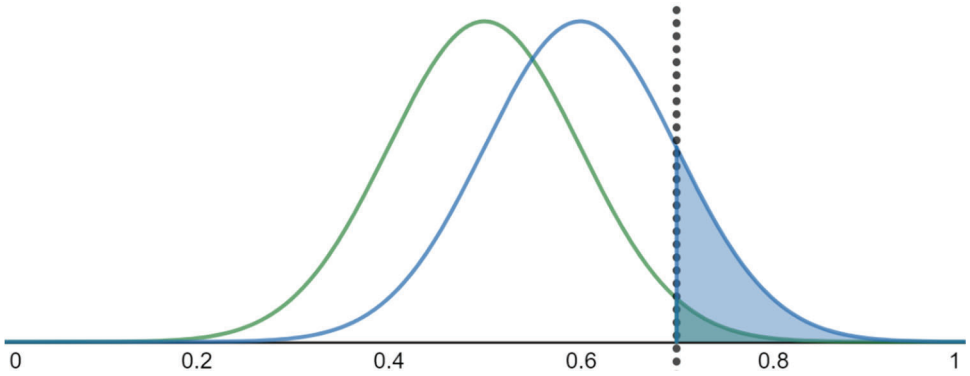
113. *See* Narayanan, *supra* note 6.

114. In the context of pretrial risk assessment, positive predictive value refers to the proportion of high-risk individuals who indeed recidivate or fail to appear in court; false negative rate is the proportion of people who would recidivate or fail to appear but were classified as low risk; and false positive rate is the proportion of people who would not recidivate and who would appear in court but were classified as high risk. *See supra* text accompanying note 6 and sources cited therein.

115. The exception is when the base rate (e.g., the probability of recidivism or failure to appear) is equal among the protected groups. *See generally* Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study Of Bias In Recidivism Prediction Instruments*, ARXIV (Feb. 28, 2017), https://arxiv.org/abs/1703.00056 [https://perma.cc/PAV8-SPT2]. Interestingly, Northpointe cited the lack of equal base rates in the observed population as a defense of its risk assessment tool during the COMPAS controversy. *See supra* text accompanying notes 48–52.

This thresholding consideration emerges in a risk assessment policy that applies a global risk threshold across local jurisdictions—including application of the same threshold to subpopulations in a given jurisdiction.[116] Suppose a county consists of two ethnic groups, Green and Blue, and adopts a risk assessment tool whose scores for the Green and Blue groups are each normally distributed with mean 0.5 and 0.6, respectively, with standard deviation 0.1, as depicted in the following figure:

Figure 1. Score distribution of Green and Blue groups: different mean.



Imagine, further, that this tool has reached some agreed-upon balance of fairness and accuracy. Distributional discrepancy within the jurisdiction can still give rise to thresholding challenges. Recall that, under SB 10, risk refers to the "likelihood that a person will not appear in court as required or the likelihood that a person will commit a new crime if the person is released before adjudication of his or her current criminal offense."[117] Applying this understanding, suppose that the statewide Council decides to categorize individuals with risk higher than 0.7 as the high risk group, indicated by the highlighted areas in Figure 1.

How might the state's global choice affect a given county and the individuals within it to whom the tool is applied? Applying the scenario described above, among the Green individuals who would neither recidivate nor miss the court date, roughly one percent will be classified as high risk. On the other hand, this number is significantly larger—ten percent—for the Blue group. That is, a Blue individual who poses no risk is ten times more likely to be classified high-risk than they would have been if they had they

116. If each sub-jurisdiction sets its own risk threshold levels by determining what quantitative scores are associated with low, medium, and high risk categories, then the understanding of "risk" would be localized, without global consensus of the sort law typically demands.

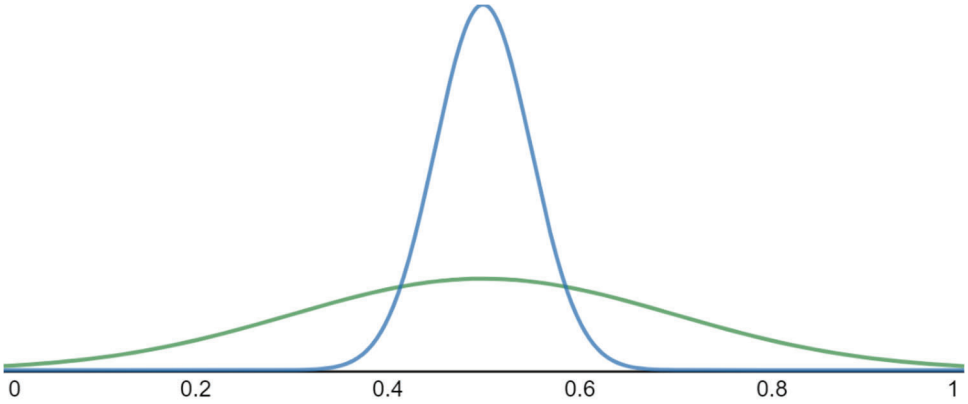
117. S.B. 10, 2017–2018 Leg., Reg. Sess. § 1320.7(h) (Cal. 2018) (enacted) (codified at CAL. GOV'T CODE § 27771 and scattered sections of CAL. PENAL CODE).

been in the Green group. Formally, the false positive rates of Green and Blue groups do not match, and the categorization would be considered unfair under this notion of fairness. In fact, in order to achieve equal false positive rates using a single threshold, we need either to classify everyone as high risk, using a threshold very close to 0, or no one as high risk, using a threshold very close to 1.

Nevertheless, we could still achieve a fair thresholding rule by setting a different threshold for each ethnic group. For instance, thresholding the Green individuals at 0.65 and Blue at 0.74 would achieve fairness with equal false positive rates of four percent. In policy terms, this would require more particularized, variable thresholding *within* a given population. But this intra-population thresholding comes at a cost: there is a tradeoff between global uniformity and localized fairness at the level of subpopulations.

Furthermore, this sort of thresholding problem can persist even when the base distributions of different subgroups are identical, due to uncertainty inherent in any statistically-derived risk assessment instrument. Consider a different hypothetical county with Green and Blue ethnic groups. Risk scores for both groups in this county are identically distributed and centered around 0.5. Suppose an algorithmic scoring system outputs a risk score with more variance for the Green group than for the Blue group. More precisely, the predicted risk scores have standard deviation 0.2 for Green and 0.05 for Blue, as shown in the below figure:

Figure 2. Score distribution of Green and Blue groups: different variance.



Such a phenomenon can often occur when one group, in this case Green, makes up a greater proportion of the population. Because algorithmic instruments are built to optimize a measure for the overall population, they will perform better for a group that makes up a greater proportion. Put differently, the tool will have a better predictive power for the majority group. Such a tool can be considered fair in the sense that the scores are

equally well-calibrated for both groups in the population as a whole. For example, any individual assigned a score 0.8 will indeed have 0.8 probability of committing a new crime or failing to appear in court. But the combination of the tool's quantitative scoring and the thresholding decision can still be considered unfair for members of minority subpopulations.

Again, applying the parameters of a statute like SB 10, suppose that a global body sets a threshold of 0.7 to classify an individual as high risk, regardless of their group membership in Green or Blue. In such a situation, about 0.81 of Green individuals in the resulting high-risk group would have recidivated or failed to show up, whereas this proportion drops to 0.71 for the Blue group. This is an example of a classifier failing to satisfy fairness as defined by predictive parity (the balance in positive predictive value between ethnic groups).[118] On the other hand, setting a threshold for each group independently could, in theory, achieve a more subgroup-sensitive understanding of fairness.[119] But this result would require far more localization of the overarching policy categories.

Thresholding of low, medium, and high levels can thus lead to unfair risk categorization in at least two ways. One, there may be unfairness due to distributional differences among subpopulations (e.g., ethnic, gender, or age groups). Two, there may be unfairness due to the inherent uncertainty in statistical risk assessment tools. In the case of a statute that sets global thresholding standards across the entire jurisdiction, as SB 10 appears to do, these sorts of thresholding issues become especially stark. In addition to underlying technical and normative questions about the "right" thresholding practices, the way that the statute or regulation allocates authority becomes critical. In particular, any global delineation might be difficult or even impossible to correct when the entities tasked with correcting unfairness are local. For instance, SB 10's draft guidance requires local courts to consider whether "any scientific research has raised questions that the particular instrument unfairly classifies offenders based on race, ethnicity, gender, or

118. Predictive parity is sometimes referred to as calibration in the technical fairness literature.

119. In fact, it has been mathematically proven that the optimal decision rule that satisfies demographic parity has to set group-specific thresholds, whether the objective is to optimize for accuracy or a more complex utility, such as balancing the social cost of releasing a high-risk individual and the cost of maintaining jails. *See* Corbett-Davies & Sharad Goel, *supra* note 105; Zachary C. Lipton et al., *Does Mitigating ML's Impact Disparity Require Treatment Disparity?*, ARXIV (Jan. 11, 2019), https://arxiv.org/pdf/1711.07076.pdf [https://perma.cc/9S3P-WHY2]; Aditya Krishna Menon & Robert C. Williamson, *The Cost of Fairness in Binary Classification*, CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY, 107, 107–18 (2018).

income level."[120] Scientific research might indeed raise questions, but whether the questions matter in fairness determinations cannot be answered in the abstract. Rather, the very definition of "unfairly classify[ing]"[121] is bound up in antecedent global choices about risk thresholds, such that this sort of global-local allocation risks asking local actors to account for choices over which they have no meaningful input or control.

This issue can, moreover, further be compounded with demographic discrepancies among counties or between counties and the state. For example, versions of Simpson's paradox mean that a threshold chosen to be fair at the state level may not achieve fairness at the level of individual counties. Recall how, in the first Green/Blue group example, the original threshold of 0.7 was adjusted down for the Green group and up for the Blue group in order to make the categorization fair in the sense of equal false positive rates. Suppose that this choice was made at the state level. If Simpson's paradox is present and the distributional difference in each county reverses the direction, such that the Green group in fact has a higher mean risk score at the county level, then this pair of thresholds set by the state would be making the risk categorization less fair at the county level. Conversely, suppose once more that every county uses the same risk threshold in an attempt to be fair across the state. This approach is likely to run into different technical issues: an attempt to threshold such that the resulting risk categorization is fair in every county may not exist unless the thresholding itself is minimal, such as, for instance, a system that assigns everyone to a single risk group. In other words, an attempt to be fair at the global level may end up being unfair at the local level, yet an attempt to be fair at the local level may be technically impossible without making the instrument close to useless. This global-local tension is a hidden consequence of the technical choices required to build risk assessment instruments.

* * *

The cumulative upshot of these sorts of technical considerations is that risk assessment tools require complex webs of policy and technical choices, globally and locally. Maximizing technical objectives will demand policy tradeoffs, and vice versa. A failure to begin with this ground-level awareness is tantamount to creating black box policy regimes that turn out to be Pandora's Boxes if we try to open them down the line. The following Part thus builds from the specific tradeoffs observed in a statute with terms like

120. Judicial Council of Cal., *Invitation to Comment: SP18-23*, CAL. CTS. 10 (2018), https://www.courts.ca.gov/documents/SP18-23.pdf [https://perma.cc/YV97-FNXQ].

121. *Id.*

SB 10 and begins to distill more generally applicable principles for the design of risk assessment statutes and regulations.

V. PATHS FORWARD

Abstracting away from SB 10's particulars, every risk assessment algorithm inevitably entails local and global allocations of authority along two related axes. One, which entity, at which level, is responsible for crafting the relevant procedures and rules or standards. This is the more traditional *policy* prong of a statute or regulation. Two, which entity, at which level, is responsible for developing, testing, and applying the instrument itself, potentially subject to local or global constraints or guidance. This is the more *technical* prong of a statute or regulation.

In practice, moreover, the picture is even more complicated because risk assessment does not allow such a crisp bifurcation of technical and policy choices.[122] Technical choices must account for local conditions to avoid unfair results, yet law's commitment to global first principles cuts against too much tailoring by jurisdiction.

This tension is especially stark for a multi-level intervention like SB 10 because the Council's proposed rules of court require each superior court and its associated PAS to ensure that the tool is "accurate," to assess whether it has been appropriately validated, and to consider whether there has been unfair classification.[123] Yet neither fairness nor the normatively proper tradeoffs between fairness and other values, like accuracy, are self-defining. Nor is there further delineation in either the statutory text or its legislative history to clarify what it means for, say, an instrument to "unfairly classif[y]" based on a sensitive characteristic.[124] The substance of these normative requirements, accordingly, will be defined locally. And given this inevitable tension between local tailoring and global commitments, clear oversight of the system itself is all the more critical to ensure that the system itself remains accountable. Too many *layers of discretion* at both the policy and technical levels risk creating a policy black box in which implementing an algorithm channels authority in unanticipated directions, potentially without

122. *Cf.* Alicia Solow-Niederman, *Administering Artificial Intelligence*, 92 S. CAL. L. REV. (forthcoming 2020) ("Algorithmic and programming decisions structure human behavior. These choices are in fact policy decisions that function at the most essential levels of democratic governance and public interests. Put simply: AI development is an especially stark example of how private coding choices are governance choices.") (on file with authors).

123. *Proposed Rules 4.10 & 4.40*, *supra* note 82, at 11–12.

124. *See id.* at 4, 11–12.

adequate democratic responsibility for the ways in which the algorithm affects actual human lives.

In the face of such complexity, we advocate simplicity. The SB 10 example suggests that risk assessment statutes will run into trouble where they create too many layers of discretion. There are two specific issues. One, if there are zones of ambiguous or even conflicting control, then there are likely to be disparate effects across localities that undermine any effort to create globally consistent policy outcomes. This will be the case when, for example, there is a top-down, statewide definition of low, medium, and high risk and a locally validated tool applied in a jurisdiction that is smaller than the state. Two, the very process of technical validation demands local determination, and an attempt to too-strictly control the tool's development and implementation top-down will undermine any effort to create risk assessment instruments that are locally accurate and unbiased. It is beyond the scope of this analysis to endorse more global or local control; such theoretical development of what to weigh in crafting a system of algorithmic governance and how to strike the right balance of global and local when it comes to both technical and policy choices awaits future research.[125] Nonetheless, simplicity counsels in favor of several preliminary lessons, with associated implications for SB 10.

Less Can be More. Tools that attempt to introduce more factors might increase accuracy, so long as the information is managed properly. Yet they might also contain more opportunities for issues, like Simpson's paradox hidden discrimination, that can elude oversight systems, particularly systems that operate at a global level and must account for many localities. For SB 10, the tools ultimately adopted, if any, should use the minimum number of factors that avoids problematic thresholding variance.

Timing Matters. Policymakers should take care in prescribing *when* global oversight is helpful, and what each global and local actor, respectively, is permitted to do at different stages of the tool development and deployment process. For example, if there is a list of globally approved tools that have been validated, as is the case for SB 10, may a locality undertake further validation to respond to a demographic change or a local policy, such as bail reform measures, that affect the likelihood of nonappearance? Risk assessment tools will make stale predictions if they are trained on historical data that does not account for more recent bail reforms.[126] To ensure that localities can update their instruments to reflect changing conditions on the ground, risk assessment statutes should both require ex post auditing of

125. *See supra* note 14 and accompanying text.

126. *See* Koepke & Robinson, *Danger Ahead*, *supra* note 13, at 24–38 (warning against "zombie predictions" that rely on stale data).

locally validated tools and be careful in calling for preapproval of tools that are removed from a particular local context.

Audit with Attention to Local and Global Detail. Several of the technical challenges—most notably Simpson's paradox—that arise in SB 10 occur because of a lack of adequate attention to local data. For example, ex post auditing at the global level alone could fail to identify local treatment of a particular subpopulation that is unfair as compared to the treatment of other demographic groups in that locality.[127] When it comes to auditing the way that a tool classifies individuals, aggregated data is not enough. Audits must take into account localized data, not merely aggregate data. Though doing so is more cost- and time-intensive, this allocation of resources must be made before deploying the tool, as part of the initial cost-benefit analysis of a turn to risk assessment algorithms.

Mixed Zones are a Mixed Blessing. Policy provisions that demand both local and global control may seem like a helpful compromise. But if they are not administered carefully, they can complicate oversight of and public accountability for risk assessment instruments. Attempts to combine local and global oversight—as seen, for instance, in SB 10's requirement that local courts assess the accuracy and discriminatory potential of tools selected from a Council-approved list—introduce a number of wrinkles.

Take validation, for instance. If a global body like the Council truly validates a tool, then it is not clear how a locality could adapt it to meet technical best practices and still permit global confidence in the tool. On the other hand, if the locality validates the tool, then substantial resources will be required for the global body to be certain that it meets its validation requirements, or there will be no meaningful oversight.

The most auspicious way to manage these global-local tensions is to approach mixed zones with caution. Caution in crafting a policy means proceeding with an ex ante awareness of and explicit delineation of which level(s) must participate at a given stage in the process, as a technical matter, and how these considerations align with the allocation of decision-making authority, as a policy matter. It also requires thinking carefully about how to grant decision-making and oversight responsibility, particularly if a proposed policy requires both local and global participation.

A Dash of Discretion. When used sparingly, allocating additional discretion within the statute might at times solve particularly thorny technical issues. For instance, in thresholding, SB 10 appears to provide that the same high/medium/low risk threshold set shall be applied to already-developed assessment tools. However, as Part IV describes, this situation could lead to

127. *See* discussion *supra* Section IV.B.

an impasse where the risk group assignment cannot be made fair without setting different thresholds for different protected subgroups or altering the tools with respect to the given threshold set. In the former instance, the global entity that sets the threshold would need additional discretion to set different qualitative risk levels associated with different subgroups, however identified. In the latter instance, localities would need additional discretion to develop and validate tools that meet the subgroup-independent thresholds in the context of that locality. Risk assessment policies, in short, must permit additional tailoring if the global thresholds are to be considered fair in particular counties, whether the discretion to tailor as required is allocated globally or locally.

Define Fairness and Specify Who Decides. Even if no single definition of fairness is likely to be without controversy, risk assessment statutes should say what they intend as a *technical* matter when it comes to such a critically contested term. When these points are unspecified, they still must be made, but the choices will tend to be implicit—as a matter of technical development—without upfront consideration, adequate opportunity for public debate, or ongoing accountability for the decision. By first clearly defining whether a global or local entity is responsible for arriving at what is fair, and, second, designing policies that designate what fairness means as a technical matter, we can better begin to grapple with the underlying normative implications of the statutory text. A critical matter for future research is whether, as a normative matter, fairness must be defined locally—lest the top-down imposition infringe on community values—or globally—lest too much tailoring to community norms contravene non-negotiable first principles.

VI. CONCLUSION

Developing and deploying risk assessment algorithms without considering how they will fit within new and existing institutions, norms, and preexisting technical and policy constraints is a mistake. The example of SB 10 highlights how risk assessment tools are not instruments that operate in isolation; rather, they are developed and deployed within legal institutions and require input from global and local decisionmakers. Control of these instruments, in turn, requires keener attention to the design of risk assessment policies, and specifically to who is granted authority and discretion over the tools. When a particular statute or regulation empowers an actor at the global level to develop a list of approved tools, for instance, how does this choice interact with the technical needs of local actors or cabin local policy discretion? Conversely, if a statute or regulation requires local validation, what limitations does this place on global oversight of the tool?

This Article encourages policymakers and technologists to ask these and related questions, by design.

Initial statutory and regulatory decisions should thus be made with attention to local-global tradeoffs, technical limitations, non-negotiable policy objectives, and underlying normative principles. A failure to grapple with these questions will not erase them. Where there are too many layers of discretion and too many local-global tensions, we would be ill-advised to rely on algorithmic risk assessment instruments as criminal *justice* tools.

STRANGE LOOPS: APPARENT VERSUS ACTUAL HUMAN INVOLVEMENT IN AUTOMATED DECISION MAKING

Kiel Brennan-Marquez,[†] *Karen Levy*[††] *& Daniel Susser*[†††]

ABSTRACT

The era of AI-based decision-making fast approaches, and anxiety is mounting about when, and why, we should keep "humans in the loop" ("HITL"). Thus far, commentary has focused primarily on two questions: whether, and when, keeping humans involved will improve the results of decision-making (making them safer or more accurate), and whether, and when, non-accuracy-related values—legitimacy, dignity, and so forth—are vindicated by the inclusion of humans in decision-making. Here, we take up a related but distinct question, which has eluded the scholarship thus far: does it matter if humans appear to be in the loop of decision-making, independent from whether they actually are? In other words, what is at stake in the disjunction between whether humans in fact have ultimate authority over decision-making versus whether humans merely seem, from the outside, to have such authority?

Our argument proceeds in four parts. First, we build our formal model, enriching the HITL question to include not only whether humans are actually in the loop of decision-making, but also whether they appear to be so. Second, we describe situations in which the actuality and appearance of HITL align: those that seem to involve human judgment and actually do, and those that seem automated and actually are. Third, we explore instances of misalignment: situations in which systems that seem to involve human judgment actually do not, and situations in which systems that hold themselves out as automated actually rely on humans operating "behind the curtain." Fourth, we examine the normative issues that result from HITL misalignment, arguing that it challenges individual decision-making about automated systems and complicates collective governance of automation.

DOI: https://doi.org/10.15779/Z385X25D2W

† Associate Professor of Law & William T. Golden Research Scholar, The University of Connecticut.

†† Assistant Professor of Information Science, Cornell University; Associated Faculty, Cornell Law School.

††† Assistant Professor of Information Sciences & Technology, and Research Associate in the Rock Ethics Institute, Penn State University. We thank Cassidy McGovern, Sherriff Balogun, and participants in the Cornell Tech Digital Life Initiative; the Cornell AI, Policy, and Practice Working Group; and the Berkeley Center for Law and Technology 2019 Symposium for helpful comments. Karen Levy gratefully acknowledges support from the John D. and Catherine T. MacArthur Foundation.

TABLE OF CONTENTS

I. INTRODUCTION

The era of automated decision making fast approaches, and anxiety is mounting about when and why we should keep "humans in the loop" (HITL).[1] Thus far, commentary has focused primarily on two questions: whether keeping humans involved will improve the results of decision making (rendering those results safer or more accurate),[2] and whether human involvement serves non-accuracy-related values like legitimacy and dignity.[3]

1. *See generally* Meg Leta Jones, *The Right to A Human in the Loop: Political Constructions of Computer Automation and Personhood*, 47 SOC. STUD. SCI. 216 (2017) (discussing the background on the burgeoning debate regarding whether to keep humans in the loop, particularly as it plays out in the United States-European Union context).

2. Medical treatment is a good example. Rich Caruana marshals a useful case study of asthmatic pneumonia patients who were categorized as "low risk" by a machine learning (ML) system—i.e., a system for automating classification tasks that infers or "learns" decision rules from prior examples rather than applying rules explicitly coded in advance—because it turns out that such patients (by contrast to non-asthmatic pneumonia patients) have historically received *much better care* from doctors, and so have displayed correspondingly better outcomes. In short, relying here on the ML system alone would have courted medical disaster. But the ML system was still a very useful input to ultimately-human decisions. *See* Rich Caruana et al., *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission*, 21 ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING PROC. 1721, 1721–25 (2015).

3. *See, e.g.*, Kiel Brennan-Marquez & Stephen Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. CRIM. L. & CRIMINOLOGY 137 (2019) (arguing that equality requires a "reversibility" dynamic between decision-makers and affected parties—and that this value

Here, we take up a related, but distinct question which has eluded the scholarship thus far: does it matter if humans *appear* to be in the loop of decision making, independent from whether they *actually* are? In other words, what is at stake in the disjunction between whether humans in fact have ultimate authority over decision making versus whether humans merely seem, from the outside, to have such authority?

Broadly speaking, our claim is that the "appearance" dimension of HITL merits exploration because when appearance and actuality are misaligned—when (1) a human appears to be in the loop, but in fact the decision-making system is fully automated, or when (2) a decision-making system appears fully automated, but is in fact bolstered by back-end human judgment—two related sets of normative issues come to the fore.

The first concerns individual experience. When appearance and actuality misalign, users of systems can become confused about what they are looking at. This dynamic risks both alienation and dignitary injuries, and deprives users of a meaningful opportunity to contest decisions.

The second set of normative issues attends to collective governance. Misalignment between the appearance and actuality of full automation can make it difficult to assess the ultimate goal of a decision-making system. Is full automation actually the desired endpoint? Are we—in the democratic, "we the people" sense—comfortable, in principle, with the automation of a given realm of decision making? Misalignment frustrates our ability to robustly ask these questions, regardless of their correct answers. Thus, where the stakes of automation are obscured by either a too-human or a falsely-inhuman veneer, democratic oversight suffers.

Our focus on the appearance of systems joins other recent legal scholarship focused on deceptive interfaces and the policy implications of humanrobot interaction.[4] Appearance emerges more latently in a good deal of other technology policy discussion. In fact, we might understand some of the most fundamental normative and policy principles in this area as efforts to align the actual and apparent operations of a system. Notice, for example, has long played a central role in policymaking around people's relationships with automated systems—most notably as a means of effectuating consent to

---

runs orthogonal to decisional outcomes); Emily Berman, *A Government of Laws and Not of Machines*, 98 B.U. L. REV. 1277 (2018) (arguing likewise); Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 GEO. WASH. L. REV. 1 (2019); Meg Leta Jones & Karen Levy, *Sporting Chances: Robot Referees and the Automation of Enforcement* (2017), https://ssrn.com/abstract=3293076 (pointing out the importance of sociocultural values like integrity and the overcoming of adversity in discussions of machine rule enforcement).

4. *See, e.g.*, Margot Kaminski et al., *Averting Robot Eyes*, 76 MD. L. REV. 983 (2017); Kate Darling, *'Who's Johnny?' Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy*, *in* ROBOT ETHICS 2.0 173 (Lin et al., eds., 2017).

data collection.[5] One of the most fundamental policy debates regarding the individualistic model of privacy regulation, and whether it can be resuscitated, involves the (in)effectiveness of privacy policies to provide notice that can serve as the basis for real consent.[6] The goal of notice, essentially, is to better align public perceptions with the actual workings of computational systems. Recent calls for interpretability of AI-driven systems, and explanations of the outcomes derived from them, have similar aims.[7]

Perhaps most fundamentally, appearances can help ensure the legitimacy of systems. Whether affected parties view decisions—particularly adverse decisions—as legitimate often depends on the presence of visible indicia of procedural regularity and fairness.[8] Sometimes, we go so far as to regulate these indicia regardless of the characteristics of the underlying system. In other words, sometimes we think appearances should be safeguarded, even if they make no difference to the ultimate decisions reached.[9] We require judicial recusal, for instance, both in cases where a judge is actually less-than-impartial, and in cases where it simply appears that way. The explicit justification for the latter—according to the American Bar Association and the Supreme Court—is that "appearance of impropriety" would "impair" the "*perception* [of a] judge's ability to carry out judicial responsibilities with integrity, impartiality and competence."[10] That is, it would threaten people's faith in the system, regardless of its impact on the case at hand.

Our argument proceeds in four parts. First, we build our formal model, enriching the HITL question to include not only whether humans are actually in the loop of decision making, but also whether they appear to be so. Second, we describe situations in which the actuality and appearance of HITL align: those that seem to involve human judgment and actually do, and those that seem automated and actually are. Third, we explore instances of misalignment: situations in which systems that seem to involve human

5. *See generally* Daniel Susser, *Notice After Notice-and-Consent: Why Privacy Disclosures are Valuable Even If Consent Frameworks Aren't*, 9 J. INFO. POL'Y 37 (2019).

6. *See, e.g.*, Ryan Calo, *Against Notice Skepticism in Privacy (and Elsewhere)*, 87 NOTRE DAME L. REV. 1027 (2012) (exploring the concept of "visceral notice" as a means of revitalizing notice-and-consent regimes).

7. *See generally* Solon Barocas & Andrew Selbst, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018).

8. *See, e.g.*, TOM TYLER, WHY PEOPLE OBEY THE LAW (2006); John W. Meyer & Brian Rowan, *Institutionalized Organizations: Formal Structure as Myth and Ceremony*, 83 AM. J. SOC. 340 (1977).

9. *See* Roger Ford, Privacy When Form Does Not Follow Function (unpublished manuscript) (on file with author) (arguing that design changes can—profitably—impact the *experience* of user interaction with technology, even if they make no difference to actual technological capacity).

10. Caperton v. A.T. Massey Coal Co., 556 U.S. 868, 888 (2009) (emphasis added).

judgment actually do not, and situations in which systems that hold themselves out as automated actually rely on humans operating "behind the curtain." Fourth, we examine the normative issues that result from HITL misalignment.

II. HUMANS ACTUALLY IN THE LOOP VS. APPARENTLY IN THE LOOP

In recent years, the HITL question has become a focal point of technology-governance scholarship. This literature offers a handful of definitions of HITL. Some commentators construe HITL narrowly—to refer, in essence, to systems that operate automatically in the mine run of cases, but that provide for human override in circumstances of obvious error.[11] Other commentators define HITL more expansively—to encompass not only the possibility of case-by-case override by humans, but also the role of humans in developing and supporting automated systems, and the co-embeddedness of humans and machines in all technology-assisted decisional environments, "automated" or otherwise.[12] Although the observation that all technical systems are socially constructed certainly has conceptual value, the observation also makes it difficult to draw meaningful lines for present purposes.

In what follows, we deploy the concept of HITL to describe any decision-making system in which the initial triage or categorization of cases is performed by a machine, but a human agent exercises some degree of meaningful influence—up to and including override—over the disposition of particular cases. Influence takes different forms. Sometimes, the human role is largely procedural: for example, pushing a given case up or back in the relevant queue, or deciding which cases merit more institutional resources. Other times, the human role is more dispositive, involving the power to shape outcomes, either in terms of a case's concrete effects (e.g., granting or denying benefits), or in terms of how the outcome is justified, or both. The specifics of the human role may vary, but the key is that a human has some form of meaningful discretion in particular cases.[13]

11. For a formal model of HITL (specifically applied to security issues, but of general relevance) that goes in this direction, see Lorrie F. Cranor, *A Framework for Reasoning About The Human in the Loop*, 1 CONF. ON USABILITY, PSY., & SECURITY PROC. (2008).

12. *See, e.g.*, Meg Leta Ambrose, *The Law and the Loop*, 2014 IEEE INT'L SYMP. ON ETHICS SCI., TECH. & ENG'G 1 (2014) (emphasizing the universality of "humans in the loop," once the category is widened to include programmers, designers, and the like).

13. Our framing here tracks the conception of humans in the loop in the discourse around the European General Data Protection Regulation (GDPR), which triggers certain protections when decisions are made "based solely on automated processing"—that is, in the absence of a human in the loop. General Data Protection Regulation, 2016 O.J. (L 119)

Further, when we talk about "particular cases," we mean instances of decision making that have a concrete impact on a specific affected party—making the dynamic of interest the triangulated interaction of (1) the automated component of the system, (2) the HITL (who gets to decide, ultimately, what the fate of the affected party will be), and (3) the affected party herself. This is a capacious definition. As a formal category, it spans a diverse array of decision-making domains, some of which involve lots of "hands-on" human involvement, others of which involve almost none. Sometimes, the HITL and the affected party may be the same person, as in decision-making systems that empower—or *seem* to empower—users to directly override machine protocols. An especially pronounced and tragic example of this arose recently in two crashes of the Boeing 737 Max, despite pilots' efforts to override the software.[14] In both cases, one could say that the pilots were both the affected party of the machine-system and the HITL—or so, at least, it appeared.

At some level, however, the key point of our HITL definition is what it does not include. It does not include human involvement in the development of decision-making systems: the human aspects of coding, product design, or supervised learning. The reason is not that such human involvement lacks normative or practical relevance in these areas. It is that we are interested primarily in the impact of HITL—in actuality as well as appearance—on specific affected parties in decisional systems.

Our primary contribution is to add a dimension to the HITL discussion. Instead of simply asking whether a human *is* in the loop, we focus on whether a human *appears to be* in the loop. In other words, what has been traditionally conceptualized as a binary question—human in the loop: "yes" or "no"—may be better conceived as a 2x2 matrix. Enriching the model in

---

(EU) (repealing Directive 95/46/EC (General Data Protection Regulation)). In discussing the meaning of this provision, the Article 29 Working Party Guidelines maintain that "fabricating human involvement"—for instance, "if someone routinely applies [machine decisions] without any actual influence on the result"—would not escape the ambit of the automated processing provision. The report further clarifies that "[t]o qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision." ARTICLE 29 DATA PROTECTION WORKING PARTY, GUIDELINES ON AUTOMATED INDIVIDUAL DECISION MAKING AND PROFILING FOR THE PURPOSES OF REGULATION 2016/679, 20-21/en. wp 251rev.01 (Feb. 6, 2018) [hereinafter GUIDELINES ON AUTOMATED INDIVIDUAL DECISION-MAKING].

14. Andrew J. Hawkins, *Deadly Boeing Crashes Raise Questions About Airplane Automation*, VERGE (Mar. 15, 2019, 1:40 PM), https://www.theverge.com/2019/3/15/18267365/boeing-737-max-8-crash-autopilot-automation [https://perma.cc/UU2L-GZ8Q].

this way—moving from a simple binary to a 2x2 matrix—helps us appreciate some of the normative complexity that attends the HITL debate.[15]

Table 1: HITL Dimensions

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop	I	II
Human does not appear to be in the loop	III	IV

On Table 1, quadrants I and IV are "aligned," meaning that the appearance of HITL and the actuality of HITL are the same. We call these quadrants *manifest humanity* and *full automation*. Quadrants II and III, by contrast, are "misaligned." Quadrant II, which we call *skeuomorphic humanity*, captures situations in which it seems like a human is present, but when a machine actually has full control. Think here of a chatbot with advanced language facility, or a home care robot that "seems human" to the patients for whom it cares. Inversely, quadrant III, which we call *faux automation*, captures situations in which the interface makes decision making seem completely automated, but where a human is actually making decisions—for example, a mobile robot that appears self-directed, but is in fact steered by a remote human driver. These definitions are included in Table 2.

15. To be sure, while our matrix adds a dimension to the HITL/no-HITL binary, it also necessarily collapses some real-life complexity. Just as a human may be more or less in the loop—that is, humans may have different degrees of discretion or autonomy vis-a-vis an automated system—the appearance of HITL is also not necessarily a binary. People may recognize, for instance, that a HITL is present, but misperceive the HITL's role. Or different users may be more or less recognizant of the true nature of the system. We elide such finer distinctions here for purposes of exploring the general dynamics, but recognize that they are likely to emerge in practice.

Table 2: HITL Dimensions with Definitions

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop	Manifest humanity	Skeuomorphic humanity
Human does not appear to be in the loop	Faux automation	Full automation

The reality, at least for the foreseeable future, is that many domains of automation will not be amenable to either of the two "aligned" quadrants. This is so for two reasons.

First, even in realms where total automation is plainly possible, the absence of humans in a process is likely to alienate some users. That is likely to inspire skeuomorphism, i.e., the *appearance* of human involvement. The companies and state agencies that develop automated technology, and the actors who deploy it, will have an incentive to use skeuomorphic techniques to drive adoption. Given this, it is plausible that many fully-automated realms will continue to maintain a veneer of human responsiveness. Techno-cultural evolution takes time.

Second, total automation will not be possible in certain realms for a long while. But it will nonetheless serve as an aspiration, and developers of technology will settle for faux automation as a bridge *toward* full automation. In other words, developers will often have an incentive to market systems which are not fully automated, on the promise—well-founded or not—that they will someday achieve full automation.

III. ALIGNMENTS

We begin with the two quadrants in which appearance and reality are consistent.

In the *manifest humanity* quadrant, a human is in fact in the loop, and this is apparent to users. Most forms of traditional adjudication fall within this category, as do uses of automated systems that serve purely to aid humans with well-established decision-making power (for example, the use of imaging technologies to assist doctors in medical diagnosis).

The inverse of manifest humanity is *full automation*—in which a process is completely and obviously automated with no human role. We may accept full automation as the best option when enforcement is low-stakes, uncontroversial, and rote—when an interest in efficiency outweighs other

normative concerns. At the other end of the spectrum, we may prefer fully automated systems in particularly high-stakes allocations of costs and benefits (like lotteries), in which we want no actual or apparent intervening value judgments about desert or blameworthiness.[16]

Each of these regimes may be advisable in some circumstances based on the values considerations we have discussed thus far (efficiency, fairness, safety, etc.). And both can be subject to legitimacy concerns on these or other grounds. We raise them here only in brief, primarily to set them aside. What interests us, ultimately, is the gap between appearance and reality—and its normative stakes.

IV. MISALIGNMENTS

A. SKEUOMORPHIC HUMANITY

Quadrant II encompasses cases of *skeuomorphic humanity*—situations in which the public generally perceives meaningful human involvement where none exists.

Human-like machine interfaces are ubiquitous. Sometimes, it is obvious to users that these machines are not actually human. Voice assistants like Siri and Alexa have notably human interactional qualities. They speak in humanoid voices, they tell jokes, and they respond to natural language queries. But their containment within a physical object like an iPhone or an Amazon Echo precludes most confusion that they are actually human. This is not always so. Online chatbots, for example, lack obvious indicia of their artificiality and often intentionally obscure it. They may do so for a variety of reasons, from efforts to deceive at scale (e.g., spambots and robocalls purporting to be from a human in need of a wire transfer) and economic and political manipulation (e.g., artificial generators of ratings and reviews; amplification of political propaganda) to therapeutic and even artistic goals (e.g., using bots to combat hate speech, or as a form of creative expression).[17] Google's artificial intelligence (AI) assistant Duplex—demonstrated at a May 2018 developer conference, in which it was used to book a haircut appointment—was purposefully given vocal qualities, tics and cadences that

16. This applies with particular force to intentionally randomized decisions. For example, Ronen Perry and Tal Zarsky discuss the attractiveness of purely random processes in high-stakes contexts like the law of the sea—if, say, one passenger must be thrown overboard to save the others, choosing the unlucky passenger by lot (presumably without subsequent appeal) may be the best way out of a bad situation. *See* Ronen Perry & Tal Z. Zarsky, *"May the Odds Be Ever in Your Favor": Lotteries in Law*, 66 ALA. L. REV. 1035, 1041 (2015).

17. *See* Madeline Lamo & Ryan Calo, *Regulating Bot Speech*, 66 UCLA L. REV. 988, 995–1002 (2019).

made it seem particularly realistic (pauses, "mm-hmm"s, and the like) to keep the person at the other end of the line from detecting its artificiality.[18]

In some cases, the skeuomorphic human is not a Siri-esque humanoid interface, but a real flesh-and-blood person—albeit one who lacks any meaningful ability to influence the relevant decision-making process. In these cases, the human is effectively no more than an ornamental aspect of the system's interface. These dynamics emerge in technical or bureaucratic systems that ostensibly involve humans, but where those humans are unable to execute discretion or diverge from administrative scripts. Think here about the familiar experience of visiting the DMV and being hamstrung by a minor technicality: for example, being told that one's insurance card needs to be in hard-copy rather than digital form in order to register a car, and that "no exceptions" can be made.[19] In practice, a human clerk is likely to deliver the news that one has failed to satisfy the agency's arcane requirements, suggesting that a well-reasoned or sufficiently emotional appeal might persuade them to revise the decision. But more often than not, the clerk merely throws up their hands and explains that they have no authority to override the rules. Although this decision-making system bears a human face, no human decision-maker impacts its outcomes (at least not in the immediate instance).

One defense of the "human gloss" is that it can make automated systems more intuitively usable. We borrow here from the vocabulary of *skeuomorphic design*—the use of design features that make an artifact resemble a previous version of itself.[20] In skeuomorphic design, the formerly functional becomes ornamental, a nod to prior technology that aids the user in transition.[21] For example, the "shutter click" sound of a phone camera: though the camera no longer has a physical shutter that makes such a sound, users have become

18. Interestingly, following blowback from critics about Duplex's deceptiveness, Google announced that a subsequent version would explicitly identify itself as an AI to the humans with whom it interacts. *See* Nick Statt, *Google Now Says Controversial AI Voice Calling System Will Identify Itself to Humans*, VERGE (May 10, 2018, 7:46 PM), https://www.theverge.com/2018/5/10/17342414/google-duplex-ai-assistant-voice-calling-identify-itself-update [https://perma.cc/5RFU-S876].

19. Readers who live in Connecticut be advised. In fairness to the state, DMV paperwork requirements were recently relaxed—registration applicants are now permitted to submit digital insurance cards. Though this, of course, does not make the system any more human; it simply makes the inhuman system more forgiving. An Act Concerning Electronic Proof of Automobile insurance identification cards, H.B. 5135, 2017 Sess. (Conn. 2009), https://www.cga.ct.gov/asp/cgabillstatus/cgabillstatus.asp?selBillType=Bill&which_year=2017&bill_num=5135 [https://perma.cc/C5EE-8NF4].

20. *Skeuomorphism*, INTERACTION DESIGN FOUND., https://www.interaction-design.org/literature/topics/skeuomorphism [https://perma.cc/29WA-JUXU] (last visited Oct. 18, 2019).

21. *See id.*

acclimated to the idea that shutter click indicates a photo taken. Therefore, subsequent technologies have included the sound as an ornament. The ornament retains social functionality by acting as a signifier, a notification to photo takers and photo subjects that a photo has been captured.[22] (Think, too, of e-readers with "pages," or digital audio controls shaped like dials.[23])

We might think of skeuomorphism as a form of *design theater*.[24] Interaction with artifacts and processes often involves a sort of ritualism; our understanding of technologies depends on how we have interacted with them in the past. When something about the technology changes in a way that obviates that ritual, we may be put off or confused. The retention of ritual—even when not strictly necessary for the system to function technically—can help the system to function socially. Consider, for instance, the legend of midcentury cake mix.[25] As the story goes, home cake mixes—in which all ingredients save water were pre-measured and mixed together, so that the baker need only dump the box's contents into water, stir, and bake—initially sold poorly. Psychologist Ernest Dichter recommended that General Mills reformulate the mix to require more human work. The reason, Dichter offered, was that housewives found the process self-indulgent: "In order to enjoy the emotional rewards of presenting a homemade cake, they had to be persuaded that they had really baked it, and such an illusion was impossible to maintain if they did virtually nothing."[26] As a result, it is said, the company changed the recipe to require that the baker add fresh eggs to the mix in place of the dehydrated eggs that had been included. This change ostensibly led to the product's wide acceptance. The story suggests that even when not essential for technical functioning, the patina of humanity in a process can matter.

Further, even in realms where we are comfortable with full automation as a normative matter—i.e., the decision-making task is not one that seems,

22. *See* John H. Blitz, *Skeuomorphs, Pottery, and Technological Change*, 117 AM. ANTHROPOLOGIST 665, 668 (2015) (describing skeuomorphs as both "utilitarian and representational"); *see also* Ivan Marković, *Vaping Like a Chimney: Skeuomorphic Assemblages and Post-Smoking Geographies*, SOC. & CULTURAL GEOGRAPHY 1, 2 (2019) (presenting a conceptual overview of the skeuomorph).

23. *See* Tim Hwang & Karen Levy, *The Presentation of Machine in Everyday Life*, WEROBOT (Mar. 2015), http://www.werobot2015.org/wp-content/uploads/2015/04/Hwang_Levy_WeRobot_2015.pdf [https://perma.cc/K6CU-TLYR].

24. *Id.*

25. The minutiae of the story itself are contested, and possibly apocryphal, but it serves its purpose here regardless. *See* David Mikkelson, *Requiring an Egg Made Instant Cake Mixes Sell?*, SNOPES (Jan. 31, 2008), https://www.snopes.com/fact-check/something-eggstra/ [https://perma.cc/8EAC-BGZL].

26. PAUL LEE TAN, ENCYCLOPEDIA OF 7700 ILLUSTRATIONS: SIGNS OF THE TIMES 1228 (1979).

in principle, to require human judgment—there may be still be dignitary reasons to maintain the appearance of humanity, even in a purely ministerial capacity. A good example is the delivery of momentous information, as in recent debates over whether doctors should deliver grave prognoses via robot.[27] Many people think that dire medical information deserves some kind of "cushion," or human gloss, which might be a freestanding argument for keeping the skeuomorphic structure in place.[28] It is also possible for the appearance of human involvement to help smoothly transition a decision-making system to full automation. This is not an argument in favor of maintaining skeuomorphic structures perpetually, but can certainly justify maintaining them in the short- to medium-term.[29] Acknowledging these benefits is quite different from wanting a human to *actually* be meaningfully involved in decision making.[30] The objection here is not to the means of arriving at the prediction, but to the method by which that prediction is communicated.

Design of this sort is not without detractors. Although some preferences are purely aesthetic, others depend on design theater's tendency to enable deception or manipulation, when users are made to feel comfortable with a new technology because they think it works just like an older one.[31] Often, design theaters operate to give users the feeling of being in greater control over a technology than they actually are (what we have elsewhere called

27. *See* David Aaro, *Family Upset After 'Robot' Doctor Informs Patient He Doesn't Have Long to Live*, FOX NEWS (Mar. 10, 2019), https://www.foxnews.com/health/family-upset-after-robot-doctor-says-patient-doesnt-have-long-to-live [https://perma.cc/M6WH-UYT7] (" 'If you're coming to tell us normal news, that's fine, but if you're coming to tell us there's no lung left and we want to put you on a morphine drip until you die, it should be done by a human being and not a machine,' Catherine Quintana told USA Today."); Evan Selinger & Arthur Caplan, *How Physicians Should and Shouldn't Talk with Dying Patients*, ONEZERO (Mar. 12, 2019), https://onezero.medium.com/how-physicians-should-and-shouldnt-talk-with-dying-patients-6ff55fcf40e4 [https://perma.cc/C39F-NS89]; Joel Zivot, *In Defense of Telling Patients They're Dying via Robot*, SLATE (Mar. 13, 2019), https://slate.com/technology/2019/03/robot-doctor-technology-patient-dying.html [https://perma.cc/8CFT-R3PP]. Notably, the human doctor did appear on the robot's screen and delivered the news via videoconference—but the means of communication nevertheless caused injury and offense.

28. *See* Zivot, *supra* note 27.

29. Katherine Metcalf et al., *Mirroring to Build Trust in Digital Assistants*, ARXIV (Apr. 2, 2019), https://arxiv.org/pdf/1904.01664.pdf [https://perma.cc/3NBJ-A8DF].

30. Brennan-Marquez & Henderson, *supra* note 3.

31. A somewhat comical, but instructive example is the "Horsey Horseless," a turn-of-the-19th-century vehicle design that consisted, essentially, of "a car with a big wooden horse head stuck on the front of it," intended to mislead *horses* on the road into accepting a motorized vehicle as one of their own. It does not appear to have worked. Alex Davies, *Well That Didn't Work: The 1899 Car With a Full-Size Wooden Horse Head Stuck to the Front*, WIRED (Feb. 10, 2015), http://www.wired.com/2015/02/well-didnt-work-1899-car-full-size-wooden-horse-head-stuck-front/ [https://perma.cc/ZU57-GFD5].

"theaters of volition")—like placebo buttons that give users the illusion of agency over elevator doors or crosswalk signals.[32] Speed is another common consideration: users may not trust computational processes that occur instantaneously, so designers may deliberately build delay and the appearance of deliberation or processing into systems.[33] Cases like these deceive users by deliberately obscuring the full capabilities of the system and the limited abilities of the human user.

Sometimes the concern is less about deception than visceral aversion. Human-like machines launch us into the uncanny valley—things that look almost, but not quite, like humans make us feel very uncomfortable.[34] There are several different explanations for this feeling of eeriness. One cognitive explanation is that when it is harder for us to categorize something immediately, we have a sense of dissonance and discomfort that is difficult to resolve. An explanation from evolutionary psychology is that vaguely unnatural movement can be an indicator of pathogens, so we are conditioned to want to stay away from it.[35] Regardless of the source, being duped by a machine masquerading as a human is an uncomfortable feeling.

More pragmatic concerns attach, too. Human-seeming systems can readily gain our trust—or manipulate us, leading to a range of consumer protection issues.[36] We may disclose more to human-seeming systems than we otherwise might, perhaps because we have misread human-like cues.[37] The mistaken sense that a human is involved in an automated process can lead people to believe that there are more opportunities for intervention and override than actually exist. Ultimately—as we explore more fully in Part IV below—the key question is whether maintaining the appearance of human involvement has sufficient benefits to outweigh the inherent shortcomings of deception.[38]

---

32. Hwang & Levy, *supra* note 23; Torin Monahan, *Built to Lie: Investigating Technologies of Deception, Surveillance, and Control*, 32 INFO. SOC'Y 229 (2016).

33. *See* Ryan W. Buell & Michael I. Norton, *The Labor Illusion: How Operational Transparency Increases Perceived Value*, 57 MGMT. SCI. 1564 (2011).

34. *See* Shensheng Weng et al., *The Uncanny Valley: Existence and Explanations*, 19 REV. GEN. PSYCHOL. 393 (2015).

35. Karl F. MacDorman et al., *Too Real for Comfort? Uncanny Responses to Computer Generated Faces*, 25 COMPUTERS HUM. BEHAV. 695, 696 (2009).

36. *See generally* Woodrow Hartzog, *Unfair and Deceptive Robots*, 74 MD. L. REV. 785 (2015).

37. Brenda Leong & Evan Selinger, *Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism*, 19 ACM FAT CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 299, 299 (2019).

38. *See generally* Eytan Adar et al., *Benevolent Deception in Human Computer Interaction*, 13 ACM CONF. ON HUMAN COMPUTER INTERACTION (CHI) (2013) (providing a thorough description of rationales and methods for user deception in human-computer interaction).

B. FAUX AUTOMATION

Quadrant III points to the inverse of skeuomorphic humanity—what is sometimes called *faux automation* (or what writer and activist Astra Taylor calls *fauxtomation*).[39] Here, the misalignment between appearance and reality arises because apparently automated systems are in fact driven by considerable human input. Of course, as scholars in science and technology studies (STS) have long argued, at some level, all technologies reflect the concerns, perspectives, and values of their human designers.[40] By *faux automation*, however, we suggest more direct forms of human involvement, consistent with our definition of HITL above.

The reason for faux automation is straightforward: building fully automated systems is hard. Despite recent advances in machine learning and AI, certain tasks that humans easily accomplish, such as understanding and using words in context, remain difficult for computers.[41] Rather than wait for further breakthroughs, technologists increasingly conceive of automation problems outside binary, all-or-nothing terms (full automation or bust), and use hybrid human-machine workflows to solve complex problems. Amazon's Mechanical Turk (AMT) system, a major platform for coordinating such work, originally described itself as facilitating "artificial artificial intelligence": a simulacrum of automation, in which humans masquerade as machines that think like humans.[42]

Examples of faux automation abound, exhibiting a variety of human-machine configurations. In some arrangements, machines do most of the work and human involvement is largely limited to quality assurance. For example, it was recently revealed that Amazon's Alexa devices—voice-activated "smart assistants" advertised as using AI to answer users' questions

39. Astra Taylor, *The Automation Charade*, LOGIC (Aug. 1, 2018), https://logicmag.io/05-the-automation-charade/ [https://perma.cc/2YCJ-2MCM].

40. *See, e.g.*, Ambrose, *supra* note 12; *see generally* BATYA FRIEDMAN & DAVID G. HENDRY, VALUE-SENSITIVE DESIGN: SHAPING TECHNOLOGY WITH MORAL IMAGINATION (2019).

41. Will Knight, *AI's Language Problem*, MIT TECH. REV. (Aug. 9, 2016), https://www.technologyreview.com/s/602094/ais-language-problem [https://perma.cc/38ZZ-XJKZ].

42. Using AMT, "requestors" distribute small work assignments ("Human Intelligence Tasks," or HITs, as Amazon calls them)—e.g., identifying objects in an image or digitizing handwritten text—to a distributed, online workforce ("turkers"), who are paid per task completed. *Artificial Artificial Intelligence*, ECONOMIST (Jun. 10, 2006), https://www.economist.com/technology-quarterly/2006/06/10/artificial-artificial-intelligence?story_id=7001738 [https://perma.cc/N4KT-FW5B].

and to control other "smart home" systems—fall into this category.[43] Unbeknownst to Alexa owners, who were given the impression that the devices are fully automated, audio recordings of user prompts and queries are regularly transmitted back to Amazon, where human technicians review them in order to tweak and improve Alexa's algorithms.[44]

At the other end of the spectrum are cases in which humans do most of the thinking and machine components are largely for show. One example is the *original* Mechanical Turk, an 18th century chess-playing automaton that turned out to have a human chess player hidden inside its enclosure. These systems are designed to give the appearance of automation without the computational substance.[45] In 2015, the public learned that the Edison automated blood testing systems sold by Silicon Valley firm Theranos were just this kind of charade.[46] Theranos advertised its Edison machines as a revolutionary technology that could process hundreds of diagnostic tests using only a few drops of blood instead of the numerous vials older techniques required. But the machines did not work.[47] Rather than admit it, the company staged misleading demonstrations and falsified Food and Drug Administration tests. The company pretended that its own machines were processing the blood, when lab technicians were actually conducting the tests behind the scenes using standard industry equipment purchased from their competitors.[48]

Many faux automated systems rely on human-machine collaborations that fall somewhere between these extremes. While significant functionality is automated, humans are generally responsible for tasks such as text and image recognition. In 2017, it came to light that Expensify (an app for generating expense reports) was using human workers contracted through AMT to digitize handwritten receipts.[49] In 2018, the Center for Public Integrity exposed widespread errors in campaign finance records caused by human mislabeling of images being prepared for automated processing by a

43. Matt Day et al., *Amazon Workers Are Listening to What You Tell Alexa*, BLOOMBERG (Apr. 10, 2019), https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio [https://perma.cc/92KW-M4AD].

44. *Id.*

45. Taylor, *supra* note 39.

46. John Carreyrou, *Hot Startup Theranos Has Struggled With Its Blood-Test Technology*, WALL ST. J. (Oct. 16, 2015), https://www.wsj.com/articles/theranos-has-struggled-with-blood-tests-1444881901 [https://perma.cc/K6G7-WWLU].

47. *Id.*

48. *Id.*

49. Alison Griswold, *Expensify's "Smart" Scanning Technology Was Secretly Aided by Humans*, QUARTZ (Nov. 30, 2017), https://qz.com/1141695/startup-expensifys-smart-scanning-technology-used-humans- hired-on-amazon-mechanical-turk/ [https://perma.cc/4WR6-9TKA].

company called Captricity.[50] Even more difficult for machines than text and image recognition is judging the meaning of words and images in context. This makes human workers essential to commercial content moderation.[51] While Facebook CEO Mark Zuckerberg has promised that the company's AI tools will rid its platform of problematic content, there is little reason to believe fully automated content moderation systems are on the horizon.[52] Facebook and other social media sites, such as Twitter and YouTube, have devised elaborate rules for determining when user-generated content should be flagged or removed—systems that Kate Klonick has likened to "legal or governance systems."[53] But in many cases, machines are incapable of determining which rules apply to particular posts, or deciding when the rules need to be revised or amended.[54] Thus, armies of human reviewers are required to carry out this interpretive work.[55]

Humans may also play a significant role in seemingly autonomous robotic systems. The Kiwibot, a four-wheeled food delivery robot currently deployed for testing on the UC Berkeley campus, is actually operated by workers in Colombia who send the robots wayfinding instructions every five to ten seconds. (The arrangement, which the company calls "parallel autonomy," saves money because the humans obviate the need for sophisticated sensor systems).[56] Similarly, a Japanese firm called Mira Robotics recently announced the release of remote-controlled "robot butlers" (think Rosie from *The Jetsons*). These robots rely on a combination of AI software for basic navigation and remote human controllers for more

50. Rosie Cima, *Company Using Foreign Workers Botches U.S. Senate Campaign Finance Records*, CTR. FOR PUBLIC INTEGRITY (Sep. 5, 2018), https://publicintegrity.org/federal-politics/company-using-foreign-workers-botches-u-s-senate-campaign-finance-records/ [https://perma.cc/G6XP-G7EN].

51. Sarah T. Roberts, *Social Media's Silent Filter*, ATLANTIC (Mar. 8, 2017), https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/ [https://perma.cc/W9K4-PAV9] ("[T]here is a profound human aspect to this work.").

52. Drew Harwell, *AI Will Solve Facebook's Most Vexing Problems, Mark Zuckerberg Says. Just Don't Ask When or How.*, WASH. POST (Apr. 11, 2018), https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/ [https://perma.cc/G2FJ-2CBG].

53. *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1602 (2018).

54. *Id.* at 1635–49.

55. *Id.*

56. Carolyn Said, *Kiwibots Win Fans at UC Berkeley As They Deliver Fast Food at Slow Speeds*, S.F. CHRON. (May 26, 2019), https://www.sfchronicle.com/business/article/Kiwibots-win-fans-at-UC-Berkeley-as-they-deliver-13895867.php [https://perma.cc/583D-WXDC].

complex tasks like folding clothes and manipulating small objects.[57] Although Mira Robotics has been forthright about its robots' human control, as these kinds of devices proliferate we can expect the gap between user perceptions about the nature of these systems and the reality of their internal functioning to grow.

Faux automation and skeuomorphic humanity are not mutually exclusive: one system can exhibit both dynamics. Consider Google's Duplex service, previously described. Originally debuted as "a new technology for conducting natural conversations to carry out 'real world' tasks over the phone," the system was designed as an outward-facing AI assistant.[58] Rather than merely answer questions, it could call and schedule reservations and appointments, speaking to other people on its user's behalf.[59] In Google's initial demonstrations, Duplex did not disclose to the people it called that they were speaking to a machine—a case of skeuomorphic humanity—and skeptics quickly raised alarms about the deception involved.[60] But a more complex revelation followed: Duplex's algorithms required significant human help in order to function. Confronted by the New York Times, Google admitted that "about 25 percent of calls placed through Duplex started with a human, and that about 15 percent of those that began with an automated system had a human intervene at some point."[61] Faux automation was thus used as a stop-gap on the way to skeuomorphic humanity—a human pretending to be a machine, while the machine pretended to be a human.

The illusion of automation gives rise to at least two distinct concerns. First, there may be contexts in which we would welcome machine assistance, but balk at human help. Smart speakers are designed to record us in what was

57. James Vincent, *Robot Butlers Operated by Remote Workers are Coming to Do Your Chores*, VERGE (May 9, 2019), https://www.theverge.com/2019/5/9/18538020/home-robot-butler-telepresence-ugo-mira-robotics [https://perma.cc/9AJY-343T].

58. Yaniv Leviathan & Yossi Matias, *Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone*, GOOGLE AI BLOG (May 8, 2018), https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html [https://perma.cc/6BTD-ABVZ].

59. *Id.*

60. *See, e.g.*, Brian Feldman, *Google Duplex Makes Your Life Easier by Making It More Difficulty for Others*, N.Y. MAG. (May 10, 2018), http://nymag.com/intelligencer/2018/05/google-duplex-no-no-no-no-no-no.html [https://perma.cc/P5RN-9RPK]; Alex Hern, *Google's 'Deceitful' AI Assistant to Identify Itself as a Robot During Calls*, GUARDIAN (May 11, 2018), https://www.theguardian.com/technology/2018/may/11/google-duplex-ai-identify-itself-as-robot-during-calls [https://perma.cc/6W42-3F6K]; Natasha Lomas, *Duplex Shows Google Failing at Ethical and Creative AI Design*, TECHCRUNCH (May 10, 2018), https://techcrunch.com/2018/05/10/duplex-shows-google-failing-at-ethical-and-creative-ai-design/[https://perma.cc/F2HV-J48C].

61. Brian X. Chen & Cade Metz, *Google's Duplex Uses A.I. to Mimic Humans (Sometimes)*, N.Y. TIMES (May 22, 2019), https://www.nytimes.com/2019/05/22/technology/personaltech/ai-google-duplex.html [https://perma.cc/H9WJ-9DFL].

once called the privacy of our own homes, and Amazon markets some of its Alexa devices, such as the "Echo Spot" smart alarm clock, for installation in the bedroom.[62] Yet those comfortable with having their intimate conversations monitored by Amazon's algorithms may feel differently about having them heard by human listeners.[63] This concern has also been raised in relation to robots: if people are "deceived into thinking the robot is acting autonomously" rather than being human-controlled, they may "disclose sensitive information to the robot that they would not tell a human, not realizing that a human is hearing everything they say."[64] This was equally true in the Expensify case, discussed above. Expensify users, under the impression that machines were digitizing their receipts, were dismayed to learn that human AMT workers read and transcribed them, as receipts often contain sensitive personal information.[65]

Second, the appearance of automation can disguise the mistreatment of human workers behind the scenes.[66] Work managed through AMT is not typically well-paid. While Amazon does not provide precise wage figures, estimates suggest that "turkers" (i.e., AMT workers) earn on average only $2 per hour.[67] In addition to wage issues, the nature of the work can be distressing and damaging. Researchers and journalists have chronicled the gruesome text, images, and videos that commercial content moderators must endure in order to purge such content from our social media feeds, and the inadequate support tech companies often provide them.[68] Yet much of this

62. Tom Warren, *Amazon's Echo Spot is a Sneaky Way to Get a Camera Into Your Bedroom*, VERGE (Sep. 28, 2017), https://www.theverge.com/2017/9/28/16378472/amazons-echo-spot-camera-in-your-bedroom [https://perma.cc/W26A-PZWY].

63. Hartzog, *supra* note 36, at 794.

64. Jacqueline Kory Westlund & Cynthia Breazeal, *Deception, Secrets, Children, and Robots: What's Acceptable?*, HUM. ROBOT INTERACTION WORKSHOPS (2015).

65. Griswold, *supra* note 49.

66. Taylor, *supra* note 39.

67. Kotaro Hara et al., *A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk*, 18 ACM CONF. ON HUM. FACTORS IN COMPUTING SYS. 11 (2018) ("We estimate that 96% of workers on AMT earn below the U.S federal minimum wage. While requesters are paying $11.58/h on average, dominant requesters who post many low-wage HITs like content creation tasks are pulling down the overall wage distribution."). Additionally, Kiwibot operators also make less than $2 per hour. Said, *supra* note 56.

68. Sarah T. Roberts, *Social Media's Silent Filter*, ATLANTIC (Mar. 8, 2017), https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/ [https://perma.cc/NDS5-9ZQE]; Sarah T. Roberts, *Meet the People Who Scar Themselves to Clean Up Our Social Media Networks*, MACLEAN'S (Jun. 15, 2018), https://www.macleans.ca/opinion/meet-the-people-who-scar-themselves-to-clean-up-our-social-media-networks/ [https://perma.cc/R6V7-DHEP]; Adrian Chen, *The Human Toll of Protecting the Internet from the Worst of Humanity*, NEW YORKER (Jan. 28, 2017), https://www.newyorker.com/tech/annals-of-technology/the-human-toll-of-protecting-the-internet-from-the-worst-of-humanity [https://perma.cc/PYP4-29NU].

work is rendered invisible, because users are led to believe that these systems are fully automated.[69]

V. HOW MISALIGNMENT UNDERMINES REASONING ABOUT AUTOMATION

The misalignments described in the previous section provoke normative worry both at the individual and institutional levels. Beneath both sets of problems lies the same fundamental issue: misalignment sows confusion. It undermines our capacity to understand and reason about automated systems. For individuals, misalignment makes it difficult to contest or resist the decisions these systems deliver. For institutions, misalignment frustrates governance; it hinders the public's ability to discern and meaningfully balance the benefits and harms of automation.

These problems manifest differently in cases of skeuomorphic humanity and in cases of faux automation. In cases involving skeuomorphic humanity, individuals confronting human-seeming, but in fact fully automated systems have no real opportunity for appeal. The human acts as a bait-and-switch, palliating users' concerns without offering real recourse. Consider again the case of a DMV agent who refuses to deviate from their administrative script, even when the decision it reaches is arguably unreasonable. Set at ease by a human veneer, we expect that a human—with the apparent power to intervene or override the system's rote determination—will hear our grievances. Instead we find that resistance is futile.[70]

In cases of faux automation, by contrast, misalignment misdirects, rather than thwarts, our attempts at contesting the system's judgments. For example, if users are given the impression that content moderation on a social media platform has been fully automated, when in fact it is carried out in large part by an army of human reviewers, they are misled about the

69. *See generally* MARY GRAY & SIDDHARTH SURI, GHOST WORK: HOW TO STOP SILICON VALLEY FROM BUILDING A NEW GLOBAL UNDERCLASS (2019).

70. Ben Wagner points out that apparent-but-not-actual HITL (what he terms *quasi-automation*, and what we call skeuomorphic humanity) can frustrate the aims of legal rules, as well. Laws that aim to promote human rights with respect to algorithmic decision-making (notably, the GDPR) assume that HITLs have some measure of agency and influence; if they do not, they amount to no more than "a human fig-leaf for automated decisions" that cannot adequately safeguard rights. Ben Wagner, *Liable, But Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems*, 11 POL'Y & INTERNET 104, 118 (2019). Wagner proposes seven criteria through which to define when a human is meaningfully in the loop, as opposed to when one is simply present to "rubber-stamp" automated decisions. *Id.* at 115.

source of problems.[71] Rather than focusing indignation on the human process that caused the mistake, people tend to lodge their grievances against automation. This result verges on ironic, since genuine automation may well be a solution to the problem (depending on our diagnosis of what the problem is), rather than its cause.

Similar issues arise at the collective or institutional level. To the extent that decision-making systems are performing sub-optimally, misalignment distorts our impression of the problem. Specifically, misalignment between the appearance and reality of human control over decision making can cause certain normative dynamics to become ambiguous or insufficiently differentiated. This is unfortunate for at least two reasons. First, different dynamics, once identified, raise different governance issues. Ambiguities between dynamics therefore produce a risk of solutions that poorly fit, or even disserve, the problem at hand. Second, the question of what dynamic we are confronting—the nature of the problem—will often be a source of normative controversy in its own right. In other words, there are many circumstances in which no "right answer" exists to the question of which dynamic is afoot. Rather, the issue is essentially and irreducibly political, such that even the question of how to conceptualize the problem calls out for democratic oversight.

To get a better sense of what we mean, consider each of the following dynamics, grouped according to which form of misalignment—skeuomorphic humanity or faux automation—they reflect. In each, we consider how normative issues can emerge based on the ideal calibration, in terms of the appearance and actuality of human involvement, for a given decision-making system.

A. DYNAMICS RELATED TO SKEUOMORPHIC HUMANITY

1. The first dynamic is that, ideally, a decision-making system would both be and appear automated—but at present it appears non-automated.

71. *See* James Vincent, *AI Won't Relieve the Misery of Facebook's Human Moderators*, VERGE (Feb. 17, 2019), https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms [https://perma.cc/2VPX-E5VQ].

Table 3: Unrealized Ideal—Full Automation

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop		PRESENT STATUS QUO (skeuomorphic humanity)
Human does not appear to be in the loop		UNREALIZED IDEAL (full automation)

Here, the skeuomorphic quadrant is essentially an interim position: the problem is not that the decision-making system *is* insufficiently automated, but that it *looks* insufficiently automated. And once again the key governance issue becomes whether it is possible—and desirable—to move toward a greater appearance, or awareness, of automation. When the answer is yes, the practical question becomes how best to facilitate the transition: by what means, on what timetable, at whose cost, and the like. Proposed chatbot disclosure laws are a good example of an effort to move in this direction.[72] By requiring overt disclosure of the machine nature of a chatbot, the user is presumably not deceived into believing she is communicating with a human, and can modulate her behavior accordingly.[73]

2. The second dynamic is that, ideally, a decision-making system would neither appear to be, nor actually be fully automated, but at present it *is* automated.

72. *See, e.g.*, Jeffrey D. Neuburger & Daryn A. Grossman, *Get All of Your Bots in a Row: 2018 California Bot Disclosure Law Comes Online Soon*, NAT'L L. REV. (June 7, 2019), https://www.natlawreview.com/article/get-all-your-bots-row-2018-california-bot-disclosure-law-comes-online-soon [https://perma.cc/98X8-6M72].

73. *But see* Lamo & Calo, *supra* note 17, at 6 (noting that even if bots are revealed as bots, they "can [still] cause harm, primarily by tricking and confusing consumers. Robocallers may deny that they are automated, call targeted individuals repeatedly, and even claim to be a representative of the IRS or another powerful entity that even a tech-savvy individual might feel too anxious to hang up on").

Table 4: Unrealized Ideal—Manifest Humanity

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop	UNREALIZED IDEAL (manifest humanity) ←	PRESENT STATUS QUO (skeuomorphic humanity)
Human does not appear to be in the loop		

The governance questions on this front are straightforward in theory, but often complex in practice. In principle, the issue is simply one of putting a human "back" into the loop—a reversion to the pre-automated world. But in practice, at least two wrinkles emerge. The first is that reversion is often costly, and directly contrary to the economic interests of the actors, governmental or corporate, who spearheaded the effort toward automation in the first place. So, at a minimum, significant political will is required. The second wrinkle is that even those who agree about the need to reinsert a human in the loop will likely dispute *how* to do so. At what point(s) in the process should human oversight be installed? And what kind of oversight? And—as ever—which humans? These issues may emerge particularly when the combination of automation and deception removes some socially important friction. For instance, while bot disclosure laws require a change in the *appearance* of a chatbot, proposed anti-robocall legislation takes a different tack by banning certain types of automated calling altogether.[74] Doing so makes direct marketing much costlier for companies making these calls, and presumably realigns their incentives to do so.

3. The third dynamic is that, ideally, a decision-making system would be automated, but not seem so, making the skeuomorphic quadrant not simply an interim state, but a direct realization of the ideal.

74. Emily Birnbaum, *Dem Chair Offers Bill to Crack Down on Robocalls*, HILL (Feb. 4, 2019), https://thehill.com/policy/technology/428372-dem-introduces-bill-to-crack-down-on-robocalls [https://perma.cc/X399-UDMU].

Table 5: Realized Ideal

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop		PRESENT STATUS QUO & REALIZED IDEAL (skeuomorphic humanity)
Human does not appear to be in the loop		

A good illustration of this dynamic is a care-bot that assists ill and elderly people.[75] Assuming for argument's sake that at least some care functions are susceptible to automation, it does not follow that "full automation" is the ideal paradigm. For it may be that other, countervailing considerations—for example, the psychological benefits that come from being cared for in a human-feeling way—may counsel in favor of continued, even perpetual, skeuomorphism. Indeed, this is precisely why many skeuomorphs exist: they lubricate the transition from Technological Environment A to Technological Environment B for the human subjects who occupy, and interact within, those environments. Sometimes, this process is self-consciously temporary. Other times, it can be indefinite, particularly when the skeuomorph evolves into a comfortable feature of Technological Environment B, despite its lack of functional purpose. Think, for instance, of the persistent use of "buttons" in UX design. There is no functional reason that screen-based interfaces must include button-shaped mechanisms of navigation. Yet people seem to like them, and understand how to use them, and it is therefore conceivable that they will persist for a long time to come.

Yet even in this case—despite the status quo overlapping formally with the ideal—many second-order governance questions remain. What are the goals of the skeuomorphic mechanism and how do they potentially trade off against other goals? Having answered that question to satisfaction, what are the specific design features of the skeuomorphic mechanism that best balance these goals?

B. DYNAMICS RELATED TO FAUX AUTOMATION

The possible dynamics with respect to faux automation form a mirror-image of those just explored.

75. Don Lee, *Desperate for Workers, Aging Japan Turns to Robots for Health Care*, SEATTLE TIMES (Jul. 30, 2019), https://www.seattletimes.com/business/desperate-for-workers-aging-japan-turns-to-robots-for-health-care/ [https://perma.cc/X5FL-NVMM].

1. The first dynamic is that, ideally, a decision-making system would both seem and be fully automated—but at present only seems automated, without actually being so.

Table 6: Unrealized Ideal—Full Automation

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop		
Human does not appear to be in the loop	PRESENT STATUS QUO (faux automation) →	UNREALIZED IDEAL (full automation)

This gives rise to two interrelated governance questions: (1) whether it is possible or realistic, given existing technology, to move toward actual automation, and (2) what the drawbacks of doing so would be. In other words, as with the equivalent dynamic above, here the faux automation quadrant is an interim state. Although full automation is the ideal, the status quo involves faux automation—and the question becomes whether it is possible (and, all things considered, desirable) to move toward the former.

Certain compliance functions are likely to fall in this category. Consider the Captricity example explored above. One might plausibly argue that it would be desirable to audit office-holders' financial data via a fully automated solution. But even so, because that ideal is not yet technologically possible, it becomes a matter of obvious public concern and accountability what types of shadow adjustments are taking place—at the behest of humans—behind the scenes.[76]

Temporary "bootstrapping" of human labor into not-yet-but-hopefully-someday-automated systems can also help us begin to understand how users are likely to interact with these systems.[77] This would allow for important research on human-computer interaction that can proceed alongside technical innovations. The "Wizard of Oz" experimental method, developed in the 1980s for human factors research, similarly involves a researcher controlling a system that a research subject believes to be autonomous, typically in order to study some aspect of the system that can be examined

76. *See* Griswold, *supra* note 49 and accompanying text (discussing the Capricity case in more detail).

77. Roboticist Wendy Ju uses this term to describe the Kiwibot's human support operation. Said, *supra* note 56.

without a fully built-out system.[78] Though researchers must always be attentive to the ethical implications of deception in research, such methods also permit much more rapid learning than would otherwise be possible.[79] But faux automation seemingly on its way to full automation can also be a fraudulent overpromise, as in the Theranos case.

2. The second possible dynamic is that, ideally, a decision-making system would neither be, nor appear to be, fully automated—but at present it has the veneer of automation.

Table 7: Unrealized Ideal—Manifest Humanity

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop	UNREALIZED IDEAL (manifest humanity)	
Human does not appear to be in the loop	PRESENT STATUS QUO (faux automation)	

This dynamic gives rise to a different set of governance questions. In essence, are there benefits associated with making actually non-automated systems look and feel more automated? We suspect the answer is almost always going to be no, for at least two reasons. The first is a simple anti-deception rationale; liberal subjects are entitled to know how the world they occupy actually works. Second, in decision-making environments that involve human judgment, we almost always care about which humans are entrusted to do the judging (and whom ought to be held to account for its outcomes). By necessity, a veneer of automation shuts that inquiry down.

Here, a good example may be content moderation. One could argue that First Amendment principles not only counsel in favor of continued human involvement in decisions about what content is so offensive or otherwise harmful that it merits restriction, but also compel us to reveal that human involvement to users. Doing so is the only way to surface the reality and dignity of the human labor required to support a system and to govern appropriately around it.

78. Paul Green & Lisa Wei-Haas, *The Rapid Development of User Interfaces: Experience With the Wizard of Oz Method*, 29 HUM. FACTORS SOC'Y, 470, 470–74 (1985).

79. Westlund & Breazeal, *supra* note 64; Hartzog, *supra* note 36, at 793–96.

3. The third possible dynamic is that, ideally, a decision-making system would involve human input, but appear to be fully automated, making the faux automation quadrant itself the optimum.

Table 8: Realized Ideal

	Human is in the loop	**Human is not in the loop**
Human appears to be in the loop		
Human does not appear to be in the loop	PRESENT STATUS QUO & REALIZED IDEAL (faux automation)	

We confess to having difficulty imagining cases that might actually populate this category and include it mostly for the sake of analytic symmetry. Nevertheless, it is possible that some cases do, or will, fall into this bucket.[80] For instance—and acknowledging the relatively far-flung nature of these examples—faux automation might be appropriate in situations where human input is desired, but where the source or nature of the input needs to be obscured. By analogy, one might think of firing squads as a kind of faux automation designed to obscure the source of human input: no one can tell which member of the squad is directly responsible for the fatality (and traditionally, one of the squad's rifles is loaded with blank cartridges to further permit each individual to disclaim moral responsibility).[81] This phenomenon is also exemplified in *per curiam* opinions, a judicial practice designed to achieve somewhat similar effects. In *per curiam* opinions, the opinion is considered to be rendered by *the court*, not by any specific judge. While these are crude approximations of cases that would actually call for the kinds of faux automation discussed in this paper, they give us reason to believe such cases—where the ideal involves disavowing but nonetheless maintaining a "human hand"—might exist. So for the moment, we leave the question open.

80. *See supra* at Part V.A.3.

81. Hanny Hindi, *Take My Life, Please*, SLATE (May 5, 2006), https://slate.com/news-and-politics/2006/05/merciful-but-messy-alternatives-to-lethal-injection.html [https://perma.cc/U3UN-T293].

VI. CONCLUSION

The age of automation is upon us. As more and more traditionally-human tasks become the province of machines, questions of governance loom large. These questions will be difficult enough in settings where the status of automation is apparent. But they will become even thornier in settings where the actuality and appearance of decision-making systems are misaligned.

In sketching our taxonomy of potential dynamics produced by misalignment, we mean to raise questions rather than resolve them. Put simply, the idea is that *any* time we are confronted with faux automation or skeuomorphic humanity, there will be at least two issues on the table. First, what kind of dynamic are we dealing with—in other words, what is the desirable end state? Second, how should we proceed within the context of that dynamic?

Both questions demand public deliberation and democratic oversight. This ideal is not always borne out in practice, for many reasons: it is costly; it relies on often-scarce political will; it becomes, at times, functionally impossible. Our point is that democratic oversight always matters in principle, even when it proves difficult in practice, and that misalignment is risky in large part because it stands to undermine such oversight. In the case of skeuomorphic humanity, the worry is that we—in the sense both of individual affected parties and of the public writ large—will be lulled, by a false sense of familiarity, into passively accepting inadvisable forms of automation. In the case of faux automation, by contrast, the worry is that we (again, in both senses) will be misled about automation's promise. We will not be able to coherently assess the costs and benefits of automation when its operation seems too good to be true.

The upshot is not that skeuomorphic humanity and faux automation are always lamentable. Each may have desirable features that override concerns about deception in particular situations. But weighing the harms of deception against other context-specific values requires knowing that deception is going on in the first place. Not only is misalignment poised to sow confusion and alienation, it's also liable, perversely, to thwart the very cost-benefit inquiry required to decide whether misalignment itself is permissible.

Going forward, the question of when misalignment is permissible—and if not, what constitutes the proper remedy—will be complex and unlikely to yield easy answers. This does not make the questions intractable. It simply requires public deliberation and democratic oversight. The future of automation, including the interplay between reality and appearance, must be something we resolve together through policy—not something imposed on us.

PROCUREMENT AS POLICY: ADMINISTRATIVE PROCESS FOR MACHINE LEARNING

Deirdre K. Mulligan† *& Kenneth A. Bamberger*††

ABSTRACT

At every level of government, officials contract for technical systems that employ machine learning—systems that perform tasks without using explicit instructions, relying on patterns and inference instead. These systems frequently displace discretion previously exercised by policymakers or individual front-end government employees with an opaque logic that bears no resemblance to the reasoning processes of agency personnel. However, because agencies acquire these systems through government procurement processes, they and the public have little input into—or even knowledge about—their design or how well that design aligns with public goals and values.

This Article explains the ways that the decisions about goals, values, risk, and certainty, along with the elimination of case-by-case discretion, inherent in machine-learning system design create policies—not just once when they are designed, but over time as they adapt and change. When the adoption of these systems is governed by procurement, the policies they embed receive little or no agency or outside expertise beyond that provided by the vendor. Design decisions are left to private third-party developers. There is no public participation, no reasoned deliberation, and no factual record, which abdicates Government responsibility for policymaking.

This Article then argues for a move from a procurement mindset to policymaking mindset. When policy decisions are made through system design, processes suitable for substantive administrative determinations should be used: processes that foster deliberation reflecting both technocratic demands for reason and rationality informed by expertise, and democratic demands for public participation and political accountability. Specifically, the Article proposes administrative law as the framework to guide the adoption of machine learning governance, describing specific ways that the policy choices embedded in machine-learning system design fail the prohibition against arbitrary and capricious agency actions

---

DOI: https://doi.org/10.15779/Z38RN30793

† Associate Professor, School of Information, University of California, Berkeley; Faculty Director, Berkeley Center for Law and Technology.

†† The Rosalinde and Arthur Gilbert Foundation Professor of Law, University of California, Berkeley; Faculty Director, Berkeley Center for Law and Technology. Much appreciation to Nitin Kohli for his expert input and close reading, Margot Kaminski for her detailed comments on an earlier draft, participants at the ACM FAT* conference, members of the UC Berkeley Algorithmic Fairness and Opacity working group, and participants in the Simons Institute for the Theory of Computer Science Summer 2019 Cluster on Fairness for insightful and helpful comments and discussions; to the Berkeley Center for Law and Technology for its support of this project; and to Sanjana Parikh and Miranda Rutherford for their superb editing and research contributions to this Article. Research for this Article has been supported by generous funding from the US NSF INSPIRE SES1537324.

absent a reasoned decision-making process that both enlists the expertise necessary for reasoned deliberation about, and justification for, such choices, and makes visible the political choices being made.

Finally, this Article sketches models for machine-learning adoption processes that satisfy the prohibition against arbitrary and capricious agency actions. It explores processes by which agencies might garner technical expertise and overcome problems of system opacity, satisfying administrative law's technocratic demand for reasoned expert deliberation. It further proposes both institutional and engineering design solutions to the challenge of policymaking opacity, offering process paradigms to ensure the "political visibility" required for public input and political oversight. In doing so, it also proposes the importance of using "contestable design"—design that exposes value-laden features and parameters and provides for iterative human involvement in system evolution and deployment. Together, these institutional and design approaches further both administrative law's technocratic and democratic mandates.

TABLE OF CONTENTS

I. INTRODUCTION

The U.S. Solicitor General's 2017 arguments opposing Supreme Court review of *Loomis v. Wisconsin*,[1] a case presenting the constitutionality of the use of risk assessment software—software that uses statistical models to predict the likelihood of an individual failing to appear at trial or engaging in future criminal activity—in sentencing, may have prevailed in convincing the Justices to deny the petition for certiorari.[2] The Solicitor General conceded that one of the issues raised in the case—"the extent to which actuarial assessments considered at sentencing" may take gender into account—"is a serious constitutional question."[3] Yet he argued that Mr. Loomis's challenge to the use of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system used by the State of Wisconsin in

1. *See* Brief for the United States as Amicus Curiae, Loomis v. Wisconsin, 138 S. Ct. 2290 (2017) (No. 16-6387), https://www.scotusblog.com/wp-content/uploads/2017/05/16-6387-CVSG-Loomis-AC-Pet.pdf [https://perma.cc/L98E-8AVH]; *see also* State v. Loomis, 881 N.W.2d 749 (Wis. 2016). The Wisconsin Supreme Court case generated petition for *certiorari*. *Id.*

2. *See Order List: 582 U.S.*, SUP. CT. U.S. 5 (June 26, 2017), https://www.supremecourt.gov/orders/courtorders/062617zor_8759.pdf [https://perma.cc/X85J-PGRK].

3. Brief for the United States, *supra* note 1, at 19.

sentencing was "not a suitable vehicle" for Supreme Court review because "it is unclear *how* COMPAS accounts for gender."[4]

Yet, however persuasive this argument might have been in the context of Supreme Court case management, the implications of this concession are shocking as a matter of policy. At no time during the challenge, which was appealed all the way to the Wisconsin Supreme Court, could the courts even determine how constitutionally relevant variables were used in the system's analysis.[5] More significantly, it is unclear whether the government ever deliberated about—or was even fully aware of—the way gender was used during the procurement of this system, or its application in the sentencing over thousands of cases.[6] The state asserted that it used "the same COMPAS risk assessment on both men and women, but then compares each offender to a 'norming' group of his or her own gender."[7] In the end, however, all evidence suggests that the State of Wisconsin left the decision of how gender was to be used at the discretion of the software vendor.

---

4. *Id.*

5. This is particularly striking because regardless of how gender is used, the decision would not constitute a trivial detail, as under the Due Process Clause, a sentencing court may not consider as "aggravating" factors characteristics of the defendant "that are constitutionally impermissible or totally irrelevant to the sentencing process, such as for example race, religion, or political affiliation." Zant v. Stephens, 462 U.S. 862, 885 (1983). The Supreme Court of Wisconsin too prohibits the use of gender as a sentencing factor. *See* State v. Harris, 786 N.W.2d 409, 416 (Wis. 2010).

6. The court record does not document any evidence of such deliberation, and we could find no evidence of such deliberation elsewhere. In fact, there are indications that the state had not even adopted high level guidelines for the design of tools. SUZANNE TALLARICO ET AL., NAT'L CTR. FOR STATE COURTS, EFFECTIVE JUSTICE STRATEGIES IN WISCONSIN: A REPORT OF FINDINGS AND RECOMMENDATIONS, 122 (2012), https://www.wicourts.gov/courts/programs/docs/ejsreport.pdf [https://perma.cc/L78K-VSRT] (suggesting that draft standards developed by a national coordinating network, which require risk tools to be "equivalently predictive for racial, ethnic and gender sub-groups represented in the Drug Court population," "*could* serve as a model for standards *should the state of Wisconsin wish to develop them*") (emphasis added). It is, moreover, difficult to assess what courts are doing to consider the embedded policies in these tools, even with substantial effort. *See generally* Robert Brauneis & Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J.L. & TECH. 103, 137–38 (2018) (reporting that only one of sixteen courts provided any information about a risk assessment tool (not COMPAS) in response to public records acts, with most claiming to be exempt).

7. *Loomis*, 881 N.W.2d at 765. The Practitioner's Guide provided by Northpointe does not mention norming. Wisconsin may be referring to either what Northpointe calls "normative subgroups," which include (1) male prison/parole, (2) male jail, (3) male probation, (4) male composite, (5) female prison/parole, (6) female jail, (7) female probation, and (8) female composite. *Practitioner's Guide to COMPAS Core*, NORTHPOINTE 1, 11–12 (Mar. 19, 2015), http://www.northpointeinc.com/files/technical_documents/Practitioners-Guide-COMPAS-Core-_031915.pdf [https://perma.cc/775A-6GMH].

While deeply troubling, this phenomenon is widespread. At every level of government, officials purchase, or contract for use of, technology systems that employ machine learning—systems that perform tasks without using explicit instructions, relying on patterns and inference instead. These systems frequently displace discretion previously held by either policymakers charged with ordering that discretion, or individual front-end government employees on whose judgment governments previously relied, with an opaque logic that bears no resemblance to the bounded and rational reasoning processes of agency personnel, but rather by patterns that machines induce by observing human actions.[8]

However, research reveals that government agencies purchasing and using these systems most often have no input into—or even knowledge about—their design or how well that design aligns with public goals and values. They know nothing about the ways that the system models the phenomena it seeks to predict, the selection and curation of training data, or the use of that data—including (as in the *Loomis* case) whether and how to use data that relate to membership in a protected class. And agencies have no input into the system's analytic technique, treatment of risk or uncertainty, preferences for false positives or false negatives, or confidence thresholds. In short, governments play no role in setting important policy.

Indeed, in a recent study by Robert Brauneis and Ellen Goodman involving open records requests seeking information about six algorithmic programs used by forty-two different agencies in twenty-three states, only one jurisdiction provided the algorithm and details about its development.[9] In most instances, by contrast, agency documents revealed that they did not have access to the algorithm, the model's design, or the processes through which the algorithm was generated or adjusted.[10] Indeed, most government bodies did not even have a "record of what problems the models were supposed to address, and what the metrics of success were."[11]

Algorithmic systems generally, and those that design and sell them, are increasingly subject to criticism for inattention to context and culture, the values baked into their design, and the biases they embed.[12] Yet government

8. *See infra* Section III.B.1 (discussing decision making by machine learning systems).

9. Brauneis & Goodman, *supra* note 6, at 137 ("[O]nly one of the jurisdictions, Allegheny County, was able to furnish both the actual predictive algorithms it used (including a complete list of factors and the weight each factor is given) and substantial detail about how they were developed.").

10. *Id.*

11. *Id.* at 152.

12. *See* Mary Flanagan et al., *Embodying Values in Technology: Theory and Practice*, *in* INFORMATION TECHNOLOGY AND MORAL PHILOSOPHY 322, 322–47 (Jeroen van den Hoven & John Weckert eds., 2008) (arguing that technology can embody values by design

agencies seeking to automate tasks left to their discretion seem persistently tone deaf to the need for greater agency and public participation in shaping technology systems. Across the country there is a smattering of public efforts to assess the policies embedded in algorithmic systems, but these are exceptions. A January 2019 Request for Proposal (RFP) issued by the Program Support Center of the U.S. Department of Health and Human Services sought a contractor who could in turn coordinate the procurement of Intelligent Automation/Artificial Intelligence (IAAI) on behalf of a range of agencies.[13] In the words of the proposal, "[t]his contract is the next logical step to integrating IAAI technologies into all phases of government operations."[14] This RFP reflects the dominant mindset of agencies: It positions machine learning systems as machinery used to support some well-defined function, rather than new methods of arranging how an institution makes sense of and executes on its mission, which is often tied to an empiricist epistemology where prediction, rather than causation, is a sufficient justification for action.[15]

The marked absence of a public sector culture of algorithmic responsibility reflects a "procurement" mindset that is deeply embedded in the law of public administration. Technology systems are acquired from third-party vendors with whom government agencies enter into contracts for goods or services. Public procurement is governed by an extensive body of regulation intended to promote certain bureaucratic values—including price,

---

and developing a framework for identifying moral and political values in such technology); Deirdre K. Mulligan & Kenneth A. Bamberger, *Saving Governance-By-Design*, 106 CALIF. L. REV. 697, 708–13 (2018) (discussing the science and technology studies as well as computer science and legal literatures on "Values in Design"); Lucas D. Introna & Helen Nissenbaum, *Shaping the Web: Why the Politics of Search Engines Matters*, 16 INFO. SOC'Y 169, 169–85 (2000) (discussing biases in the creation of search indexes and search results); James H. Moor, *What is Computer Ethics?*, 16 METAPHILOSOPHY 266, 266–75 (1985) (discussing the ethical implications of invisible abuse, emergent bias due to designers' values, and bias rooted in complexity within computer systems).

13. *See* Aaron Boyd, *HHS Contract Will Offer AI Tech, Support to All of Government*, NEXTGOV.COM (Jan. 10, 2019), https://www.nextgov.com/emerging-tech/2019/01/hhs-contract-will-offer-ai-tech-support-all-government/154078/ [https://perma.cc/W8NH-CYHY].

14. *Solicitation/Contract/Order for Commercial Items: Solicitation Number 19-233-SOL-00098*, U.S. DEP'T HEALTH & HUM. SERVS. 9 (Jan. 10, 2019) https://www.fbo.gov/utils/view?id=39d0a0ce8bfe09391b9fee07833274de [https://perma.cc/6DEC-L5WQ] [hereinafter *Solicitation Number 19-233-SOL-00098*].

15. Rob Kitchin, *Big Data, New Epistemologies and Paradigm Shifts*, BIG DATA & SOC'Y 3–5 (2014), https://doi.org/10.1177/2053951714528481 [https://perma.cc/3N7Q-3LYG] (describing and critiquing Big Data "empiricism, wherein the volume of data, accompanied by techniques that can reveal their inherent truth, enables data to speak for themselves free of theory").

fairness in the bidding process, innovation, and competition[16]—and elaborates methods of challenging contracting decisions on these elements. This body of regulation generally limits standing to challenge contracting decisions to jilted commercial competitors. Both public contracting and decision making about agency management are largely exempted from administrative procedures that govern decisions of policy[17]—procedures intended to promote a different set of public values: substantive expertise, transparency, participation and political oversight, and reasoned decision making. Thus, current agency perception and practice leave the policies that algorithms embed obscured, unarticulated, and unvetted.

This Article makes the case that because choices in the design, adoption, and use of machine learning systems often make substantive policy, design, adoption, and use should be approached with a different mindset—a "policymaking" mindset—and should reflect the frameworks for legitimate policymaking embodied in administrative law.

Designing algorithmic and machine learning systems involves decisions about goals, values, risk and certainty, and a choice to place constraints on future agency discretion. If these systems employ adaptive machine learning capabilities, their design choices make policy—not just once when they are designed, but over time as they adapt and change. When the adoption of those systems is governed by procurement, the policies they embed receive little or no agency or outside expertise beyond that provided by the vendor: no public participation, no reasoned deliberation, and no factual record. Design decisions are left to private third-party developers. Government responsibility for policymaking is abdicated.

An important body of scholarship has explored the possibilities and shortcomings inherent in algorithmic systems,[18] suggested ways in which

16. *See generally* Steven L. Schooner, *Desiderata: Objectives for a System of Government Contract Law*, 11 PUB. PROCUREMENT L. REV. 103 (2002) (summarizing nine goals identified for government procurement systems: competition, integrity, transparency, efficiency, customer satisfaction, best value, wealth distribution, risk avoidance, and uniformity).

17. *See, e.g.*, 5 U.S.C. §§ 553(a)(2)–(3) (2012) (containing the Administrative Procedure Act's exemption of matters relating to "agency management" or to "public property, loans, grants, benefits, or contracts" from the section's general requirements of notice-and-comment rulemaking).

18. *See, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015) [hereinafter BLACK BOX]; Jane Bambauer & Tal Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV. 1 (2018); Kenneth A. Bamberger, *Technologies of Compliance: Risk and Regulation in a Digital Age*, 88 TEX. L. REV. 669, 724 (2010); Peter A. Winn, *Judicial Information Management in an Electronic Age: Old Standards, New Challenges*, 3 FED. CTS. L. REV. 135 (2009); Guy Stuart, *Databases, Felons, and Voting: Bias and Partisanship of the Florida Felons List in the 2000 Elections*, 119 POL. SCI. Q. 453 (2004); Kate Crawford, *The Hidden Biases in Big Data*, HARV. BUS. REV. (Apr. 1, 2013), https://hbr.org/

individual government determinations based on algorithmic systems might be challenged,[19] and proposed methods for increasing transparency and accountability.[20] Fewer researchers have extended these insights to accommodate the pressing challenges of machine learning,[21] and even fewer have explored what moving technology systems acquisition and design from a "procurement" mindset to a "policymaking" mindset would mean in terms of technical design, administrative process, participation, and deliberation.[22]

This Article begins to fill that gap. It argues that, in contexts in which policy decisions are likely to be made through procurement, process suitable

---

2013/04/the-hidden-biases-in-big-data [https://perma.cc/MH6U-28M2]; Julia Angwin et al., *Machine Bias: There's Software Used Across the County to Predict Future Criminals. And it's Biased Against Blacks*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/WK73-BW9S].

19. Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1252 (2008).

20. *See* Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 633 (2017) (suggesting a "technological toolkit to verify that automated decisions comply with key standards of legal fairness"); Frank Pasquale, *Restoring Transparency to Automated Authority*, 9 J. TELECOMM. & HIGH TECH. L. 235, 235–36 (2011); Katherine Fink, *Opening the Government's Black Boxes: Freedom of Information and Algorithmic Accountability*, INFO., COMM. & SOC'Y 1–19 (May 30, 2017), https://doi.org/10.1080/1369118X.2017.1330418 [https://perma.cc/ATP4-KRZ8] (reviewing current state of law and practice with respect to whether algorithms would be considered "records" under the Freedom of Information Act and reviewing agency bases for withholding algorithms and source code under FOIA requests); *see also* Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 121 (2019) (arguing that recently introduced provisions protecting employees against trade secret actions could immunize whistleblowers policing algorithms from within firms).

21. *See, e.g.*, Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1 (2019); Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147 (2017); Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399 (2017).

22. Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 6, 26–30 (2019) (proposing a regulatory toolkit to govern the use of algorithms in the private sector, including "substantive rulemaking mechanisms, such as the use of safe harbors and private sector codes of conduct, and accountability mechanisms, such as the use of oversight boards and audits"); Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 110 (2017) (calling for "criminal justice expertise and political process accountability" to be brought into the design of recidivism risk tools); Andrew D. Selbst, *Disparate impact in big data policing*, 52 GA. L. REV. 109, 109 (2017) (recommending police be required to complete "algorithmic impact statements" before adopting predictive policing technology); Catherine Crump, *Surveillance Policy Making By Procurement*, 90 WASH. L. REV. 1595 (2016) (proposing steps to strengthen democratic input); Dillon Reisman et al., *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*, AI NOW INST. (Apr. 2018), https://ainowinstitute.org/aiareport2018.pdf [https://perma.cc/N6W6-JRHQ]. Danielle Citron's work examining a prior generation of expert systems has provided foundational analysis for thinking about ways that administrative law concerns about delegation and process might be translated to the technological context. Citron, *supra* note 19, at 1252.

for substantive administrative determinations should be used: process ensuring the type of deliberation that safeguards fundamental administrative law values. Such processes must satisfy administrative law's *technocratic* demands that policy decisions be the product of reasoned justifications informed by expertise—elements grounded in the rule of law.[23] And they must reflect *democratic* requirements of public involvement and political accountability. The Article thus makes the case that the policies designed into machine learning systems adopted by government agencies must be surfaced and deliberated about through new processes and brought fully within an administrative law mindset. Governance through technology cannot be allowed to quietly route around the processes that ground agency action's legitimacy.

Part II describes the ways that the integration of machine learning into governance has been viewed as a matter of procurement and the failures of that approach. Government agencies have relied on private vendors for the design of algorithmic systems, largely exacerbating the challenges of governing through technology by abdicating government's role in shaping important design choices. It then explores five examples of ways in which system design embeds policy decisions to make the case that machine-learning system adoption should often instead be understood as policymaking.

Part III examines administrative law as an alternative framework for the adoption of machine learning in governance. Describing the specific ways in which machine learning systems displace administrative discretion and human logic, this Part argues that the policy choices embedded in system design fail the prohibition against arbitrary and capricious agency actions absent a reasoned decision-making process that enlists the expertise necessary for reasoned deliberation, provides justifications for such choices, makes visible the political choices being made, and permits iterative human oversight and input. This Part focuses on changing the system-adoption process, arguing that design choices should occur through a decision-making process that reflects the technocratic and democratic goals of administrative law.

Finally, Part IV envisions what models for machine learning adoption processes that satisfy the prohibition against arbitrary and capricious actions might look like. It first explores processes by which agencies might overcome the problems of system opacity and their own lack of technical expertise, satisfying administrative law's technocratic demand for reasoned expert

23. Kevin M. Stack, *An Administrative Jurisprudence: The Rule of Law in the Administrative State*, 115 COLUM. L. REV. 1985, 1989 (2015) (grounding reasoned justification as a rule-of-law requirement).

deliberation. Specifically, we urge the reliance on centers of expertise—on the model of the U.S. Digital Services Team (USDS) and the 18F "skunk works" team first developed by the Obama Administration—that develop and provide shared technical knowledge in ways that address expertise gaps across agencies, while providing a systemic approach to the use of technology in government activity.

Part IV explores both institutional and engineering design solutions to the challenge of policymaking opacity, offering process paradigms to ensure the "political visibility" required for public input and political oversight as well as proposing the importance of using "contestable design." Contestable systems foster user engagement by exposing value-laden features and parameters, and provide for iterative human involvement in system evolution and deployment in a way that would foster agency staff's awareness and participation as policies embedded in systems evolve dynamically. Together, these institutional and design approaches further administrative law's democratic mandate.

Where machine learning systems "learn" and "exercise discretion" in ways that are not guided by reasoned human decision-making inputs, and then make substantive policy that alters the legal rights and responsibilities of individuals, policymaking fails the touchstone obligation that agency actions not be "arbitrary and capricious." It shirks the requirement that decisions reflect reason, facts, context, and the factors mandated by Congress in the relevant organic statute, while avoiding elements extraneous to the legislative command.[24] The current adoption of such systems through procurement processes threatens the very premises for administrative delegation to agencies and deference to their decisions, such as expertise, reasoning, flexibility, and accountability. We outline a path to realign these powerful new tools with democratic ideals.

II. THE PROCUREMENT MINDSET: A MISMATCH FOR MACHINE LEARNING ADOPTION

A. THE ALGORITHMIC TURN IN GOVERNANCE

The State of Wisconsin's decision to purchase the COMPAS system from the Northpointe Corporation reflects an accelerating public administration trend. The increasing availability of machine-learning products and services has ushered in reliance on algorithmic decision-support and decision-making systems throughout all levels of government.[25] Agencies increasingly recognize the promise and power of systems employing artificial

24. Motor Vehicle Mfrs. Assn. v. State Farm Mut., 463 U.S. 29, 42–43 (1983).
25. Mulligan & Bamberger, *supra* note 12.

intelligence and machine learning for augmenting human administrative capacity. On the one hand, they more accurately and effectively analyze and learn from data while identifying and managing risk;[26] on the other, they "reduc[e] repetitive administrative tasks," thus freeing government "employees to focus their time and human capacity on higher value activities and decisions."[27]

Artificial intelligence and algorithmic systems are being employed, on the one hand, to better automate processes like data management and procurement, and on the other, to automate decision making across a range of substantive contexts. Federal agencies have employed technology systems for tasks ranging from determining the level of different veterans' disabilities for purposes of compensation[28] to identifying fraud in a variety of public benefits programs.[29] States and localities rely on a wide range of analytic systems "to generate predictive models to guide the allocation of public services";[30] to govern individual determinations such as teacher evaluations,

26. Coglianese & Lehr, *Transparency*, *supra* note 21, at 6 (describing how privately developed algorithms produce "unparalleled accuracy" compared to other statistical methods and human judgment).

27. *Solicitation Number 19-233-SOL-00098*, *supra* note 14, at 7 (quoting OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, OMB BULL. NO. M-18-23, SHIFTING FROM LOW-VALUE TO HIGH-VALUE WORK (2018)).

28. *See* Bob Brewin, *Goodbye paper: VA Installs Automated Claims System in All Regional Offices*, NEXTGOV.COM (Jan. 10, 2019), http://www.nextgov.com/health/2013/06/goodbye-paper-va-installs-automated-claims-system-all-regional-offices/65030/ [https://perma.cc/7D9D-JC6P]; Marion-Florentino Cuéllar, *The Surprising Use of Automation by Regulatory Agencies*, REG. REV. (Dec. 20, 2016), https://www.theregreview.org/2016/12/20/cuellar-surprising-use-of-automation-agencies/ [https://perma.cc/HUL3-Y6NJ] ("[T]he software took over this responsibility for determining levels of disability from Department 'raters'—human beings charged with determining a claimant's entitlements.").

29. *See* Leon Erlanger, *The Tech HHS, SEC, SSA and Other Agencies Use to Ferret Out Cheaters and Crooks*, FEDTECHMAGAZINE.COM (May 10, 2017), https://fedtechmagazine.com/article/2017/05/tech-hhs-sec-ssa-and-other-agencies-use-ferret-out-cheaters-and-crooks [https://perma.cc/QC55-7HGA] (discussing the ALERT program, intended to identify suspicious transactions under the SNAP ("Supplemental Nutrition Assistance Program") food stamps program, and the Social Security Administration's application to disability benefits); *see also* Coglianese & Lehr, *Regulating by Robot*, *supra* note 21, at 1162–67 (discussing initiatives within the Post Office, the Environmental Protection Agency, the Internal Revenue Service, the Federal Aviation Administration, and the Food and Drug Administration).

30. Brauneis & Goodman, *supra* note 6, at 107; *see* Coglianese & Lehr, *Regulating by Robot*, *supra* note 21, at 1161 (providing examples); 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 (2017) (discussing different ways the Internal Revenue Service uses data mining to solve the problem of tax noncompliance).

bonuses, and terminations;[31] and to identify the risk that children are victims of abuse or neglect.[32] Advocates have recently determined that local police departments are relying on Amazon Web Services facial-recognition product Rekognition to assist in identifying suspects,[33] and the FBI has announced its intention to do so as well.[34] Axon, a company producing body cameras, plans to introduce real-time facial recognition software into the products it provides to law enforcement;[35] the data obtained would be used by police departments, but retained and analyzed by Axon on Axon's own cloud services.[36] The Department of Homeland Security, moreover, has proposed procurement for an "extreme vetting" machine learning system that seeks to make "determinations via automation" as to whether an individual seeking a visa for entry to the United States will be a "positively contributing member of society," will "contribute to the national interests," or "intends to commit criminal or terrorist acts."[37]

Recent work by regulation scholars Cary Coglianese and David Lehr has drawn an important roadmap of the ways in which machine learning's capacity might enable widespread application within the administrative state.

---

31. *See* Marissa Cummings, *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/JA6A-G3MW] (discussing artificial intelligence system used by the City of Houston, the Education Value-Added Assessment System (EVAAS), which made decisions about teacher evaluations, bonuses, and terminations based on variables including student's performance on prior standardized tests); *see also Settlement Agreement*, AFT.ORG (Oct. 2, 2017), https://www.aft.org/sites/default/files/settlementagreement_houston_100717.pdf [https://perma.cc/SC7W-XZEC] (agreeing that teachers would no longer be terminated based primarily on their EVAAS score).

32. Allegheny County Department of Human Services, *Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions*, ALLEGHENY COUNTY ANALYTICS (May 1, 2019), https://www.alleghenycountyanalytics.us/index.php/2019/05/01/developing-predictive-risk-models-support-child-maltreatment-hotline-screening-decisions/ [https://perma.cc/YPX4-J3P8].

33. Nick Wingate, *Amazon Pushes Facial Recognition to Police. Critics See Surveillance Risk*, N.Y. TIMES (May 22, 2018), https://www.nytimes.com/2018/05/22/technology/amazon-facial-recognition.html [https://perma.cc/WSF9-Q7ZK] (discussing facial recognition use in Orlando and Washington State).

34. Frank Konkel, *The FBI is Trying Amazon's Facial-Recognition Software*, NEXTGOV.COM (Jan. 3, 2019), https://www.nextgov.com/emerging-tech/2019/01/fbi-trying-amazons-facial-recognition-software/153888/ [https://perma.cc/BZL7-EX6T].

35. Ian Wren & Scott Simon, *Body Camera Maker Weighs Adding Facial Recognition Technology*, NPR (May 12, 2018), https://www.npr.org/2018/05/12/610632088/what-artificial-intelligence-can-do-for-local-cops [https://perma.cc/4JUQ-NCFK].

36. *Id.*

37. *Extreme Vetting Initiative: Statement of Objectives (SOO)*, FEDBIZOPPS.GOV (June 12, 2017), https://www.fbo.gov/utils/view?id=533b20bf028d2289633d786dc45822f1 [https://perma.cc/UZQ9-9N8W].

In the future they envision such systems will permit the conduct of adjudications "by algorithm" and rulemakings "by robot" without any human involvement, facilitating, for example, full automation of the antitrust review process, or real-time dynamic evolution of the Securities and Exchange Commission's rules governing market transactions.[38] Regardless of whether this future is desirable, it emphasizes the broad and deep policy implications of these technical systems.

B. CHALLENGES OF ALGORITHMIC GOVERNANCE: VALUES IN TECHNOLOGY DESIGN

An extensive body of research suggests the difficulties inherent in attempts to use technology systems in government decision making. As a basic matter, translating legal values into design requirements is difficult. While legal policy "tempers rule-based mandates with context-specific judgment that allows for interpretive flexibility and ongoing dispute about the appropriateness of rules, [some] computer code operates by means of on-off rules,"[39] and all software systems and models require up-front decisions about what data they can assess. The social and technical environment, in which regulatory norms and the norms of coders who actually design technology are "translated" into code, exacerbates divergences between "law in the books" and "law in emerging technology."[40] Thus, even if technology accurately captures intended legal variables in its design, it may not clearly reflect the choice of how to respond to outputs "in a normative sense."[41] As a result, depending on who is involved in the process of translation, technical solutions for enabling, enforcing, or restricting rights and values can result in unintended consequences—consequences that privilege certain stakeholders and values at the expense of others.[42]

---

38. Coglianese & Lehr, *Regulating by Robot*, *supra* note 21, at 1171–84.

39. Mulligan & Bamberger, *supra* note 12, at 710.

40. Mireille Hildebrandt & Bert-Jaap Koops, *D7.9: A Vision of Ambient Law*, FUTURE IDENTITY INFO. SOC'Y 22 (Oct. 4, 2007), http://www.fidis.net/fileadmin/fidis/deliverables/fidis-wp7-d7.9_A_Vision_of_Ambient_Law.pdf [https://perma.cc/T47H-Y3CV].

41. Noëmi Manders-Huits, *What Values in Design? The Challenge of Incorporating Moral Values into Design*, 17 SCI. ENG. ETHICS 271, 279 (2011) (describing what she calls "The Naturalistic Fallacy").

42. *See* Alvin M. Weinberg, *Can Technology Replace Social Engineering?*, *in* TECH. & FUTURE 28, 34 (Albert H. Teich ed., 11th ed. 2009); Andrea Roth, *Trial by Machine*, 104 GEO. L.J. 1245, 1252–69 (2016) (describing how the processes of developing and adopting technical systems in the criminal justice system, largely driven by law enforcement, produced a hyper focus on the elimination of false negatives); Eaglin, *supra* note 22, at 101–04 (describing how normative policy judgments are delegated to the developers of actuarial risk assessment tools whose incentives and preferences can produce tools in conflict with public laws and policies); *see also* Carsten Orwat & Roland Bless, *Values and Networks—Steps Toward Exploring*

Moreover, algorithmic decision-making systems are biased. They make classification decisions based on selected data that may be inadequate or unrepresentative, improperly cleaned or interpreted, and reflect historical and ongoing structures of discrimination.[43] For example, machine learning algorithms trained from human-tagged data inadvertently learn to reflect biases of the human taggers.[44] Two years ago, academics conducted studies showing that human annotators of data used in systems exhibit core human biases that end up decreasing the accuracy of the system at large.[45] On a more general scale, Lisa Gitelman in her book *Raw Data is an Oxymoron* notes that no data is free from certain bias since it is all "cooked" at some point by software, whether it be by end-users on an online platform or by back-end algorithms.[46] Even if the training data is adequate, tagged correctly, and minimizes inherent bias, predictive algorithms can still insert inaccuracy into a given system. Predictive algorithms are essentially autonomous profiling by a machine-learning system.[47] While the aim of predictive algorithms is to identify correlations and make predictions about behavior at the individual level, the system uses groups or profiles to do so. In some systems, these groups may be constantly changing as the algorithm identifies more salient patterns. This redefinition sometimes creates profiling algorithms that correlate bias in outputs.[48] The system's algorithms in turn learn from such data, generating their own biases through the features they identify and the weights they place on them.

These concerns are paramount in machine learning systems that learn and adapt while in use. Such systems blur the line between implementation and policymaking.[49] To the extent technical systems generally are perceived as mere tools straightforwardly implementing policy choices determined elsewhere, the use of systems that change over time surely cannot fit within

---

Their Relationships, 46 COMPUTER COMM. REV. 25, 28 (2016) (discussing how technology choices can shift "costs or other burdens to parties not involved in decisionmaking").

43. Tal Z. Zarsky, *Transparent Predictions*, U. ILL. L. REV. 4 (2013); Nicholas Diakopoulos, *Algorithmic Accountability: Journalistic Investigation of Computational Power Structures*, 3 DIGITAL JOURNALISM 398 (2015); Rachel Courtland, *Bias Detectives: The Researchers Striving to Make Algorithms Fair*, NATURE (2018).

44. Diakopoulos, *supra* note 43, at 398.

45. Ishan Misra et al., *Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels*, 2016 IEEE CONF. ON COMPUTER VISION & PATTERN RECOGNITION (CVPR) 2930, 2930 (2016).

46. LISA GITELMAN, RAW DATA IS AN OXYMORON 2 (2013).

47. Mireille Hildebrandt & Bert-Jaap Koops, *The Challenge of Ambient Law and Legal Protection in the Profiling Era*, 73 MOD. L. REV. 428 (2010).

48. Tal Z. Zarsky, *supra* note 43, at 4.

49. Citron makes a related but distinct point that automated systems "blur the line between adjudication and rulemaking, confounding the procedural protections governing both systems." Citron, *supra* note 19, at 1278.

this fiction. In this context especially, designing for values is complicated, as the data and models, as well as the policy implications, shift at design, configuration, and run time.[50]

These fundamental concerns—about whether system design reflects desired values, embeds bias, or produces inaccurate results—are exacerbated by the opacity of algorithmic systems. Scholars have identified a number of ways that these systems operate as black boxes, inscrutable from the outside:[51] (1) corporate secrecy, by which the design details are kept secret by private developers;[52] (2) technical illiteracy—the impenetrable nature of system rules to non-engineers even where they are shared; and (3) the inability of humans, even those who design and deploy machine learning systems, to understand the dynamic models learned by complex machine learning systems.

Each of these levels of opacity plague government agencies seeking to employ machine learning in governance, which most often lack the technical expertise to design or assess algorithmic systems on their own. The resulting concerns are aggravated in the context of algorithmic systems used for public governance—rather than in the private sector—as the innards of the software system that privatizes public functions are typically shielded from public scrutiny.[53]

Government procurement procedures are often extensive and time consuming. They are focused on promoting management goals, such as competition, integrity, transparency, efficiency, customer satisfaction, best

50. David D. Clark et al., *Tussle in Cyberspace: Defining Tomorrow's Internet*, 13 IEEE/ACM TRANSACTIONS NETWORKING 462, 466 (2005).

51. Jenna Burrell, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, BIG DATA & SOC'Y 3.1 (2016).

52. *See* Brauneis & Goodman, *supra* note 6, at 38–44 (reporting on cities' use of trade secrecy to limit responses to Public Record Act requests for information about algorithms); *id.* at 44–47 (reporting on cities' resisting Public Record Act requests about algorithms due to concerns about gaming or circumvention and other concerns); Rebecca Wexler, *When a Computer Program Keeps You In Jail*, N.Y. TIMES (June 13, 2017), https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html [https://perma.cc/47VB-R2JU] (discussing trade secrecy limitations on access to the algorithms used in the COMPAS system at issue in the *Loomis* case); Danielle Keats Citron, *Open Code Governance*, 2008 U. CHI. L. F. 355, 357 (2008) ("Because these systems' software is proprietary, the source code—the programmers' instructions to the computer—is secret.").

53. Fink, *supra* note 20, at 1–19 (reviewing current state of law and practice with respect to whether algorithms would be considered "records" under the Freedom of Information Act (FOIA), and reviewing agency bases for withholding algorithms and source code under FOIA requests and finding exemptions claimed under national security, privacy, law enforcement investigations as well as trade secrecy exemptions).

value, wealth distribution, risk avoidance, and uniformity.[54] They are not focused on restricting the privatization of public functions, or providing oversight over delegated policymaking. Recent attempts to promote agency technology modernization by focusing on empowering Chief Information Officers in the approval, certification, and ongoing oversight of IT systems[55] have not addressed the need for agency participation in system design. In many instances, the technology is a commercial off-the-shelf product purchased after a period of market research and a general solicitation.[56] So agencies use procurement processes to acquire these complex technical systems much as they do to purchase other goods, despite these systems' widespread implications for governance and policy.[57]

This "procurement mindset" reflects a number of phenomena. As an initial matter, much of the adoption of machine learning systems through procurement no doubt comes from the perspective that these systems simply supply a new process or practice for fulfilling the agency's mission. Indeed, some such systems look more administrative in nature.

Second, procurement of off-the-shelf products often reflects a realistic assessment of agency capacity, in light of the opacity and complexity of algorithmic systems.[58] On the one hand, private developers keep much of the relevant code secret. On the other hand, agency staff frequently have few technical skills, so they can neither assess technology design shared with them nor participate in design themselves.

54. Schooner, *supra* note 16, at 104. They are, moreover, notoriously burdensome and slow, especially in the context of a dynamic information technology landscape. The Federal Acquisition Regulation (FAR), 48 C.F.R. §§ 1–53, for example, which sets forth the federal procurement process, establishes best practices, procedures, and requirements for agencies, and provides standard clauses and forms. 48 C.F.R. §§ 52–53. In addition, the FAR expressly authorizes agency heads to issue agency-specific procurement regulations implementing or supplementing the FAR, meaning that agency procurement varies greatly from agency to agency. 48 C.F.R. § 1.3; *see* FEDERAL CHIEF INFORMATION OFFICER COUNSEL, STATE OF FEDERAL INFORMATION TECHNOLOGY 36 (2017).

55. Federal Information Technology Acquisition Reform Act § 101(a), 40 U.S.C. § 11319 (2012); OFFICE OF PERSONNEL MGMT., OFFICE OF THE CHIEF INFO. OFFICER, FITARA COMMON BASELINE IMPLEMENTATION PLAN: FISCAL YEAR 2016 11 (2016), https://www.opm.gov/about-us/open-government/digital-government-strategy/fitara/opm-fitara-common-baseline-implementation-plan.pdf [https://perma.cc/4HG3-FC64].

56. *See* 48 C.F.R. § 14.101 (setting forth the "negotiated contract" process that would generally be used in the federal system for acquiring complex software involving machine-learning algorithms).

57. *See, e.g.*, Houser & Sanders, *supra* note 30, at 865–66 (2017) (discussing findings that underlying databases used by IRS algorithms "seriously" lacked supporting documentation, implicating their accuracy).

58. Burrell, *supra* note 51.

Finally, even if government bodies realize that there are important decisions embedded in systems, agencies may believe that those decisions do not constitute "policy" in the way that law traditionally understands it. Under the Federal Administrative Procedure Act, for example, matters related to agency management and contracts are both exempt from procedures that govern the adoption of policy.[59] Accordingly, the Internal Revenue Service (IRS), one of the few agencies to publicly address agency participation in system design—and one that relies heavily on algorithmic systems—has stated publicly that its use of "decision analytics" and "data and predictive modeling" constitutes "internal enforcement policy" that does not require public feedback during its development.[60]

C. EXAMPLES: POLICY IN SYSTEM DESIGN

Decisions about how to design these systems[61]—as well as how they are configured and how agency staff interact with them—touch on, and at times embed decisions about, traditional substantive policy questions. This Section identifies five of these types of determinations[62] and provides illustrative examples in which abdicating policy questions has led to real-world failures.

1. Optimization Embeds Policy

The choice of task for which a machine learning system is designed to optimize rests on, among other things, a set of assumptions about human behavior and social structure. Using such systems to govern implicates not only questions about whether the assumptions are well-founded generally, but also about how widely-applicable they are: do they reflect all individuals,

59. *See* 5 U.S.C. § 553(a)(2) (2012) (excepting such matters from the requirements of informal rulemaking).

60. TAXPAYER ADVOCATE SERV., 2010 ANNUAL REPORT TO CONGRESS: IRS POLICY IMPLEMENTATION THROUGH SYSTEMS PROGRAMMING LACKS TRANSPARENCY AND PRECLUDES ADEQUATE REVIEW 80 (2010), https://www.irs.gov/pub/irs-utl/2010arcmsp5_policythruprogramming.pdf [https://perma.cc/3PDD-D369] [hereinafter TAXPAYER ADVOCATE SERV.].

61. Several scholars have explicated the design processes of computer systems to reveal a broader range of interventions to address legal and policy concerns. *See* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 669–700 (2017) (arguing that machine learning's "playing-with-the-data stages" demands attention from legal scholars as the stages of problem definition, data collection, data cleaning, adjustment based on summary statistics, decisions about the portion of a data set to use for training versus testing, and model selection and training present both distinct opportunities and risks to accuracy, explainability, and discrimination, among others); Kroll, *supra* note 20 (describing how computational methods can be used throughout the design of computer systems generally to ensure procedural regularity).

62. This is an illustrative, not exhaustive, set of examples. Each of the "playing-with-the-data stages" described by Lehr and Ohm require choices that, depending upon the substantive context, may embed what administrative law would consider substantive policies.

groups, or situations rather than just some? These assumptions also determine the factors that will be most salient in the choices made by algorithmic systems—the governance metrics—in much the same way traditional policy decisions identify which factors should guide any administrative decision.

The governing power of assumptions is reflected in an algorithmic system used by the United States Department of Agriculture (USDA) in the food stamp program. The system used transaction records created by the "electronic benefit transfer," or debit cards, issued by the government to monitor stores for evidence of fraud. In 2002, based on a determination by that system, the agency disqualified several grocery stores serving predominantly Muslim East African communities from accepting federal food stamps—a decision with significant effect, as these stores supported the religious dietary needs of families that relied on the federal program.[63]

Citing the system's conclusion, the agency informed the relevant shops that "a careful analysis revealed . . . transactions that establish a clear and repetitive pattern of unusual, irregular and/or inexplicable activity for your type of firm," and on that basis eliminated them from the program.[64] The suspicious transactions at issue included large purchases made minutes apart, transactions for even-dollar amounts—described as unusual for food purchases—and instances in which a few households made several unusually large purchases in a single month, which is, the USDA letter stated, "not consistent with the conditions in your store for store physical size, stock of eligible items, lack of counter space and the lack of carts and baskets."[65]

The model that the system deployed to identify fraud rests on a particular assumption: fraud will manifest in certain behaviors shared across groups. Yet consideration of demographically specific spending patterns would have identified that assumption's flaws. Culture affects food purchasing habits in profound ways, rendering a one-size-fits-all model inappropriate. The purchasing patterns identified as anomalous may be normal for a subset of the population. Rather than being an indication of fraud, the patterns reflect how religion, nationality, economics, food preparation, and ordering behavior influence the purchasing behavior of a specific community.

63. Florangela Davila, *USDA disqualifies three Somalian markets from accepting federal food stamps*, SEATTLE TIMES (Apr. 10, 2002), http://community.seattletimes.nwsource.com/archive/?date=20020410&slug=somalis10m [https://perma.cc/Q2GZ-W6BZ].

64. Chris Mcgann, *Somali Merchant Waits And Hopes*, SEATTLE POST-INTELLIGENCER REP. (July 1, 2002), https://www.seattlepi.com/news/article/Somali-merchant-waits-and-hopes-1090433.php [https://perma.cc/6KXW-W8Z9].

65. *Id.*

In fact, as reporter Florangela Davila explained in an analysis of the USDA's action,[66] East African immigrant women often shop in groups of two or three, which explains why transactions from the same household often occur in pairs and threes. Because East African immigrants often lack transportation, they tend to make large consecutive purchases in fewer shopping trips. It is customary to make larger purchases of Halal meat in one trip to the market, and even to buy an entire goat, spending as much as $150 at a time, to be frozen and eaten over weeks or a month. And a habit of ordering meat by the dollar amount, rather than the pound, produces the supposedly anomalous large number of even-dollar purchases at the relevant stores.[67]

As this episode demonstrates, because algorithms optimize over large sets of data, distinct patterns in small subpopulations are obscured by design. The decision of whether to generate one model to identify fraud across all users of electronic benefits, or separate models that attend to variations in subpopulations, is a policy decision.

The recidivism risk system at issue in *Loomis* also reflects the ways assumptions set policy. Loomis's expert witness set out several questions along these lines related to the design of the COMPAS assessment. He noted that the "Court does not know how the COMPAS compares that individual's history with the population that it's comparing them with. The Court doesn't even know whether that population is a Wisconsin population, a New York population, [or] a California population," and argued that "it is critical that it be validated for use in the jurisdiction that is planning to use it."[68] In fact, a report provided to the State of Wisconsin had emphasized the need for such local validation, but the state had not performed one prior to using the tool.[69]

A California study recommending rejection of the same COMPAS system used by Wisconsin discussed the problem of validating for relevant

66. Davila, *supra* note 63.

67. *Id.*

68. State v. Loomis, 881 N.W.2d 749, 756–57 (Wis. 2016).

69. At the time this tool was used to make decisions about Loomis, Wisconsin had not undertaken a local validation, despite determining its necessity. *Loomis*, 881 N.W.2d at 762 ("Wisconsin has not yet completed a statistical validation study of COMPAS for a Wisconsin population."); Suzanne Tallarico et al., *supra* note 6, at 22–23. While the Wisconsin Supreme Court allowed COMPAS risk assessments to be used at sentencing, it circumscribed its use by requiring Presentencing Investigation Reports that contained them to inform the sentencing court that the "risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed" along with information about the secrecy of the model, the need for monitoring and updating, and research raising concerns about disproportionate classification of minority offenders as high-risk. *Loomis*, 881 N.W.2d at 764. Whether judges understand the implications of the lack of local validation is unknown.

populations extensively, asking: "Will the results generalize to other samples?"[70] The study concluded that "it is unclear" whether the formulas will generalize from this New York sample of probationers to other samples of offenders.[71] To the extent that the predictors of recidivism differ across groups, these formulas may not work in some of the California Department of Corrections and Rehabilitation's (CDCR) primary populations of interest (e.g., inmates, parolees). Given how actuarial formulas are derived and issues of over-fitting, it is necessary to cross-validate actuarial formulas with a sample of individuals from the population of interest.[72]

These examples illustrate the importance of identifying the fault lines between populations for a given task. As in the assistance fraud case above, relevant subpopulations may not always be evident up front but rather only be discovered using exploratory machine learning approaches. Further, even if subpopulations are identified, the question remains whether or not they should or may be subject to different models. These decisions regarding whether and how to segment populations for different models is a core question of policy.

2. Decisions About Target Variables Embed Policy

In social science, a key element of research design is identifying how to construct an experiment that will test the phenomena of interest. The term "construct validity" is used to ask whether the observations—choices of both instrumentation and data—will actually capture the phenomena of interest. In machine learning systems, the same question also arises. For example, when creating a risk assessment tool, one must determine how to operationalize the risk of recidivism.[73] The decision about how to optimize the target variable has sweeping policy implications.

In the recidivism risk context, an agency might like to measure actual recidivism but lacks the data to do so: we simply do not have ground truth to know whether any individual will commit a crime after release. Because recidivism itself cannot be measured, re-arrest is used as an outcome variable in the model.

---

70. Jennifer L. Skeem & Jennifer Eno Loudon, *Assessment of Evidence on the Quality of the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)*, U.C. IRVINE, at 22–23 (2007), http://ucicorrections.seweb.uci.edu/files/2013/06/CDCR-Skeem-Eno Louden-COMPASeval-SECONDREVISION-final-Dec-28-07.pdf [https://perma.cc/EMT7-7CJY] (prepared for the California Department of Corrections and Rehabilitation).

71. *Id.*

72. *Id.*

73. *Id.* at 5 (discussing construct validity: "it must measure the criminogenic needs it purports to measure; for example, it should relate coherently to other measures of needs and capture change in risk state over time").

As many have pointed out, however, arrests are both poor and biased proxies for actual recidivism rates.[74] First, incomplete observations mean we do not know all outcomes. Second, re-arrest rates reflect policing patterns, which historically police communities of color at higher rates than white communities. Thus optimizing for arrests in place of recidivism creates systems that overrepresent populations in ways that play past discrimination forward *by design*.[75] A study conducted for the CDCR rejected a recidivism risk tool on these grounds, stating that there is no evidence "that it assesses the criminogenic needs it purports to assess"[76] and concluding that the tool "reliably assess(es) something that looks like criminogenic needs and recidivism risk" but "there is little evidence that this is what . . . [it] actually assesses."[77]

Given that many of the phenomena we use models to measure—such as risk—cannot be truly observed, the proxies we select to measure them reflect policy choices about how best to measure and predict the phenomena.

3. The Choice of Model Embeds Policy

Designers of machine learning systems must choose a modeling framework.[78] Consequently, the choice of framework embeds a theory of how or why a phenomenon is occurring in the world.[79] For example,

74. Eaglin, *supra* note 22, at 94–95 (discussing inherent bias in selecting re-arrest as the measure of recidivism, given disproportionate police scrutiny of minority communities); Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, 46 CRIM. JUST. & BEHAV. 185, 196–97 (2019) (describing research that finds "people of color, especially Black people, are more likely to be arrested than Whites for the exact same behavior," which makes arrest a racially tainted proxy for recidivism).

75. Eckhouse, *supra* note 74, at 197. Eckhouse stated:

> When both the data used to produce the risk-assessment instrument and the data used to evaluate it come from the criminal justice system, quantitative risk assessments merely launder that bias. . . . [T]he legitimating process of quantitative assessment converts unequal data-generating processes into apparently objective data, without removing the fundamental problems.

Id.

76. Skeem & Loudon, *supra* note 70, at 6.

77. *Id.*

78. *See* Lehr & Ohm, *supra* note 61, at 688–95 (distinguishing between different general classes of models—random forests, neural networks, etc.—highlighting six considerations that can influence model selection, and explaining that models can be chosen from pre-configured options bundled into software and services, or be modifications made to an existing model).

79. This Section benefited from the analysis of Nitin Kohli. Memo and conversation are on file with author.

PredPol,[80] the predictive policing system adopted off the shelf, and without public process, by dozens of police departments across the United States (including Los Angeles and the University of California, Berkeley),[81] uses a "seismological" model to describe how crimes propagate throughout a region.[82] The motivation for using a seismological approach is that an original crime has ripple effects that lead to other crimes, much as an earthquake can lead to aftershocks that propagate through space and time. This model, known as Epidemic Type Aftershock Sequence (ETAS), decomposes crime into two components: background events and aftershock events. If the model is valid, then this decomposition will allow the system to predict when and where similar crimes are likely to occur, given information of recent criminal activity.

The ETAS model transplants assumptions valid in scientific geological models to the criminal context. In doing so, it implicitly constructs a particular theory of how crimes ripple through time and space. Yet the approach threatens to trade off modeling simplicity with real-world boundary conditions: the behavior of earthquakes cannot accurately predict key elements in modeling background and aftershock effects of crime, such as whether all violations have the same value—by producing the same sort of aftershocks—or whether the kind or location of crime factors into the modeling. In the seismological context, moreover, it is appropriate to assume that aftershock effects are less common after longer periods of time, less common at locations far away from the source, and uni-directional—radiating outward. But aftershock crimes need not obey the same constraints.

The assumptions embedded in the model become more problematic in conjunction with the known limitations of data about crime discussed above. Historical crime data is a lower bound on the actual representation of crime, raising important issues of selection bias and generalizability. More importantly, any bias in crime reporting patterns—for example, social stigmas related to reporting some crimes and risks of reporting that vary by context—further reduces PredPol's knowledge about "aftershock events" of the observed crime, let alone all crime in general. Simply put, the social laws

80. PREDPOL, https://www.predpol.com/ [https://perma.cc/Q9N4-9CSP] (last visited Sept. 30, 2019).

81. Caroline Haskins, *Dozens of Cities Have Secretly Experimented With Predictive Policing Software*, VICE (Feb. 6, 2019, 7:00 AM), https://www.vice.com/en_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software [https://perma.cc/N384-4HJW].

82. The analysis of the PredPol algorithm that follows is based on information provided on their website as well as academic papers (also referenced by PredPol) to provide insight into the domain specific methods. *Predictive Policing: Guidance on Where and When to Patrol*, PREDPOL, https://www.predpol.com/how-predictive-policing-works/ [https://perma.cc/5LCY-6N25] (last visited Sept. 30, 2019).

that govern the spread and reporting of crime do not obey the physical laws that govern the spread and visibility of earthquakes.

Choosing a model is a significant decision of policy. Doing it well requires an understanding of how the model relates to relevant domain- or field-specific theories of the phenomena of interest, a careful examination of any properties a model inherits from its domain of initial development, and an examination of the way model choices might introduce bias.

4. Choosing Data on Which to Train a Model Embeds Policy

Machine learning systems generate algorithms based on sample data, known as "training data," in order to make predictions or decisions without being explicitly programmed to perform those tasks. They are then shaped through feedback gleaned by the system's observations of additional data. Thus, the choice of the data on which to train a model will have profound implications for the model's outputs.[83]

An analysis of the COMPAS system provides an excellent example of the rigorous way in which data must be interrogated to determine if they are likely to produce an accurate model for a given population. The authors of the report explain:

> [T]he COMPAS data are not representative of the California Department of Corrections and Rehabilitation inmates because among other things, eight groups of inmates with potentially greater needs, including those with mental health classifications and those targeted for the substance abuse programs, were excluded from the sample . . . [moreover,] it is not clear how these offenders compare to offenders in other states. Moreover, the data are largely based on offenders' self-report, and there is no protection against reporting bias, including exaggeration or minimization of needs.[84]

The selection and use of protected attributes like race and gender within a dataset used to train machine learning models is a particularly significant policy decision. While it may be that some uses of gender could advance justice, that does not mean that such use would survive an equal protection

83. Many scholars have offered excellent explanations and examples of data-related bias; these four are particularly rich and powerful: Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 677–94 (2016); 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/WC3X-CBJD]; and ROB KITCHIN, *Conceptualising Data*, *in* THE DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES & THEIR CONSEQUENCES 1–26 (2014).

84. Skeem & Loudon, *supra* note 70, at 24 (citing JEFFREY LIN, PAROLEE NEEDS IN CALIFORNIA: A DESCRIPTIVE ANALYSIS OF 2006 COMPAS DATA, THE CENTER FOR EVIDENCE-BASED CORRECTIONS (2007)).

clause challenge. As the U.S. Supreme Court has said, the "Equal Protection Clause [is] not to be rendered inapplicable by statistically measured but loose-fitting generalities."[85] Even where the use of gender may serve a just purpose—as the State of Wisconsin claimed in *Loomis*—the Court has upheld disparate treatment based on gender only where it seeks to level the playing field. As the Court established in *Mississippi University for Women v. Hogan*: "In limited circumstances, a gender-based classification favoring one sex can be justified if it intentionally and directly assists members of the sex that is disproportionately burdened."[86] The proposition that including gender as a factor might enhance the predictive accuracy of a model or that using it to normalize results improves predictive accuracy for both men and women however, is, on its own, a legally insufficient reason for choosing whether or how to use it. This use cuts to the heart of equal protection law; as the Court notes, "proving broad sociological propositions by statistics is a dubious business, and one that inevitably is in tension with the normative philosophy that underlies the Equal Protection Clause."[87]

Thus, the question of how protected attributes are used in a statistical model, whether with pen and paper or by software algorithm, is a question of great political and legal importance. The COMPAS manual describes a system that uses gender in several ways. It presents sixteen "common categories or prototypical offending and behavior patterns that often reappear in criminal justice populations" for use in treatment planning which are segmented along gender lines.[88] It provides users with the ability to consider scale scores in reference to the scale distributions of eight normative subgroups that again are broken down along gender lines.[89] These choices about how to use data cut to the heart of commitments to equal protection and are surely substantive policy.

5. Decisions About Human-System Interactions Embed Policy

Last but not least, the interfaces and policies that structure interactions between agency staff and machine learning systems shape policy outcomes. The ways humans and machines are bound together through interfaces,

85. Craig v. Boren, 429 U.S. 190, 209 (1976).

86. Miss. Univ. for Women v. Hogan, 458 U.S. 718, 728 (1982).

87. *Craig*, 429 U.S. at 204.

88. NORTHPOINTE, *supra* note 7, at 48–49.

89. *Id.* at 11 (the current normative subgroups for comparison are "(1) male prison/parole, (2) male jail, (3) male probation, (4) male composite, (5) female prison/parole, (6) female jail, (7) female probation, and (8) female composite").

processes, and policies in "automation policy knots"[90] shape their impact. A few examples illustrate their importance.

First, as noted above, local police departments, and now the FBI, are relying on Amazon Rekognition to assist in identifying suspects.[91] Like many other software products, Rekognition has preconfigured defaults. The default "confidence threshold" for the face-matching is 80%. Leaving the default confidence threshold as such, ACLU researchers found that it incorrectly matched twenty-eight members of Congress with arrestees in the database—a 5% error rate among legislators—with a disproportionate number of false positives for African-American and Latino members. While Amazon's system documentation contains some language recommending law enforcement to use a confidence threshold of 99%,[92] the out-of-the-box default does not appear to have any particular relation or relevance to the domains in which it is being used. More importantly, the choice of threshold implicitly makes a policy decision about the tradeoffs between false positive and false negatives.[93] Such choices are paradigmatic questions of policy—they do not have answers in data but reflect instead value judgments that

90. Meg Leta Jones, *The Ironies of Automation Law: Tying Policy Knots with Fair Automation Practices Principles*, 18 VAND. J. ENT. & TECH. L. 77 (2015) (revealing how legal approaches that ignore the complex relations between humans and machines fail to protect the values they were drafted to protect); Steven J. Jackson et al., *The Policy Knot: Re-integrating Policy, Practice and Design in CSCW Studies of Social Computing*, PROC. CSCW '14 1 (Feb. 15, 2014) (coining the term policy knot: "practices and design impact and are impacted by structures and processes in the realm of policy").

91. Wingate, *supra* note 33.

92. *Amazon Rekognition Developers Guide*, AMAZON 143 (2019), https://docs.aws.amazon.com/rekognition/latest/dg/rekognition-dg.pdf [https://perma.cc/P8TH-RL5W]. The guide states:

> All machine learning systems are probabilistic. You should use your judgment in setting the right similarity threshold, depending on your use case. For example, if you're looking to build a photos app to identify similar-looking family members, you might choose a lower threshold (such as 80%). On the other hand, for many law enforcement use cases, we recommend using a high threshold value of 99% or above to reduce accidental misidentification.

Id.

93. *See* Eckhouse, *supra* note 74, at 194–95 (describing the lack of guidance on how to translate a risk score produced by a recidivism risk tool into categories to support judges or other decision-makers, and the way that categorizations can inflate perceptions of risk if they deviate from mental models about how a five-part scale, for example, would relate to overall risk); *see also* Eaglin, *supra* note 22, at 85–87 (discussing how translation between numerical output of algorithmic system and risk is subjective policy choice).

should reflect the goals set for the agency, refined through policy processes, not the designer's preference.[94]

A second example comes again from the recidivism risk domain and involves the policies governing decision maker behavior in the face of system determinations. Specifically, jurisdictions can establish policies that make departures from the recommendation of a recidivism risk system procedurally and potentially politically costly, placing a thumb on the scale of outcomes. For example, a judge may be required to justify a decision to deviate from a risk score on the record.[95] Such provisions not only raise the time required to exercise discretion but also may make judges vulnerable when they release an individual who later commits a crime, against the system's recommendation, thus contributing to a culture of deference to machine reasoning even when the law might prefer human judgment. Or as in the *Loomis* case, judges may be burdened with detailed information about the system's construction—its lack of local validation, for example—but lack the expertise to understand their practical meaning. This information mandate places a human more in the loop but it is unclear how this particular actor—a legal professional—at this particular juncture can protect the values of due process and equality put at risk by ill-chosen design.[96]

Therefore consequential decisions about people's lives are delegated in a variety of ways to machine learning systems that governments buy, or more typically contract to use—generally in an off-the-shelf manner. In effect, governments are outsourcing decisions of policy—sometimes life-changing ones—to algorithmic systems with little understanding of the assumptions those systems embed, the logics on which they rely, the data on which they were trained, or any of the other information necessary to understand whether or not they adequately and appropriately perform the reasoning tasks being handed to them.

Moreover, in the most shocking instances, institutions within the justice system have procured and used recidivism risk systems without understanding the embedded definitions of fairness, the confidence thresholds, the limitations presented by choices of training data sets, or the

94. *See, e.g.*, Felicitas Kraemer et al., *Is There an Ethics of Algorithms?*, 13 ETHICS & INFO. TECH. 251 (2011) (describing the ways that value judgments regarding false positives and false negatives govern the choice between different rational design decisions, and the setting of thresholds); Eaglin, *supra* note 22, at 88.

95. *See, e.g.*, N.J. STAT. ANN. § 2A:162–23 (West 2017).

96. *See* Jones, *supra* note 90, at 90–100 (describing how laws designed to achieve a goal by removing or inserting a human in the loop without thoroughly considering how the knot of policies, processes, and design work in practice—taking a socio-technical systems view—often fail, and advocating a set of Fair Automation Practice Principles to guide the construction of human-machine collaborations).

systems' use of protected characteristics. Individuals whose lives are being altered by these black box decision-making systems are not the only ones who do not understand how these systems work. In an unprecedented dereliction of oversight, government agencies at all levels are, perhaps unwittingly, turning over key policy questions to privately developed algorithmic systems.[97]

At various points over the past fifty years, policymakers have recognized the substantive nature of decisions that can be masked by procurement, and have suggested alternative models to ensure administrative processes of the type usually accorded traditional types of policy decisions. In 1969, for example, the Administrative Council of the United States recommended that, consistent with the goal of "assur[ing] that Federal agencies will have the benefit of the information and opinion that can be supplied by persons whom regulations will affect," the exemption from notice-and-comment rulemaking procedures for matters relating to "public property, loans, grants, benefits, or contracts" be discontinued,[98] and several federal agencies, at different points in time, required such procedures for procurement decisions.[99] More recently the IRS Taxpayer Advocate advocated (unsuccessfully) for subjection of IRS "policy guidance embedded in [automated] systems"[100]—which are neither reviewed internally nor published—to the "same stringent vetting and review process as written instructions or policies."[101] Those written policies undergo a formal

97. Mulligan & Bamberger, *supra* note 12, at 741 ("Public power is too often exercised in private, by private parties, or without nonpartisan or nonpolitical sources of expertise. The substance and political nature of choices fixed by technology is thus obscured, which enfeebles citizen awareness and involvement, diminishes ex post accountability, and yields unintended outcomes.").

98. Admin. Conference of the U.S., Recommendation number: 69-8, Elimination of Certain Exemptions from the APA Rulemaking Requirements (Oct. 22, 1969); 38 Fed. Reg. 19784 (July 23, 1973).

99. *See, e.g.*, 36 Fed. Reg. 13804 (July 24, 1971); Revocation of Statement of Policy on Public Participation in Rulemaking, 78 Fed. Reg. 64194 (Oct. 28, 2013) https://www.federalregister.gov/documents/2013/10/28/2013-25321/revocation-of-statement-of-policy-on-public-participation-in-rulemaking [https://perma.cc/7ZTG-NWHM] (showing the Department of Agriculture's history of forty-two years of notices and comments); 29 C.F.R. § 2.7 (1979) (promulgated at 36 Fed. Reg. 12976 (July 10, 1971)). The C.F.R. section states:

> It is the policy of the Secretary of Labor that in applying the rule making provisions of the APA the exemption therein for rules relating to public property, loans, grants, benefits or contracts shall not be relied upon as a reason for not complying with the notice and public participation requirements thereof.

Id.

100. TAXPAYER ADVOCATE SERV., *supra* note 60.

101. *Id.* at 78.

clearance, subject to public scrutiny, by staff of the Taxpayer Advocate Service, who review proposed guidance for conflicts with existing policies and procedures, for technical accuracy, and to identify policies or procedures that may harm taxpayers, and offer solutions and alternatives to alleviate these burdens.

Each of these experiments suggests a shifting mindset for structuring the adoption of algorithmic systems—from procurement to administrative process. The next Part takes up these suggestions; those aspects of machine learning systems that touch on substantive aspects of the relationship between the citizen and the state must be viewed as policy and should be brought within the framework that maintains and constrains the exercise of agency power.

III. BRINGING MACHINE-LEARNING SYSTEM DESIGN WITHIN ADMINISTRATIVE LAW

A. ADMINISTRATIVE PROCESS FOR MACHINE LEARNING DESIGN

Identifying the ways that the design of machine learning systems can embed value decisions reveals the ways that the adoption of machine learning systems through procurement can render policymaking invisible. Design choices set policy without input from agency employees, stakeholders, or other experts. The models, assumptions, metrics, and, at times, even the data that drive such systems, are largely opaque and unknown to government officials who acquire them and the public they govern.

When such systems embed policies, the current method of adoption lacks all hallmarks of legitimate governance. Administrative actors are excused from reasoning, analysis, and the requirement that they justify policy choices. They bring no expertise to bear. They elicit no public participation or input. Their decisions evade judicial review and political oversight. Scholarship has largely failed to address this phenomenon of lawless governance. To be sure, a robust literature has focused on the challenge of system opacity, proposing algorithmic "transparency" as a means to address the ways opacity can obscure bias, error, and outcomes that diverge from public goals.[102] Proposals for transparency have focused on open sourcing a

102. *See generally* Charles Vincent & Jean Camp, *Looking to the Internet for Models of Governance*, 6 ETHICS & INFO. TECH. 161, 161 (2004) (explaining that automated processes remove transparency); Paul Schwartz, *Data Processing and Government Administration: The Failure of the American Legal Response to the Computer*, 43 HASTINGS L.J. 1321, 1343–74 (1992) (setting forth an influential paradigm for addressing data-driven governance, which includes making data processing systems transparent; granting limited procedural and substantive rights to the data subject; and creating independent governmental monitoring of data processing systems).

given system's software code and releasing it to the public for inspection; mandating disclosure of system methodology;[103] disclosing the sources of any data used;[104] requiring audit trails that record the facts and rules supporting administrative decisions when they are based on automated systems; mandating that hearing officials explain in detail their reliance on an automated system's decision;[105] and notifying those affected when algorithmic systems are used.[106]

Such transparency mechanisms, in turn, are intended to facilitate accountability.[107] Openness about the algorithms that drive technological systems government agencies use permits public analysis and critique,[108] and an assessment of the fairness of their use. It allows software audits[109] that identify correct, and incorrect, inputs and outputs, back-testing of those input and outputs to assure the system is executing its intended goals,[110] and testing of software on specific scenarios with pre-determined outcomes. It can allow individuals to contest, inspect, and adjudicate problems with data or decisions made by a system, facilitating challenges to government determinations based on algorithmic systems. Such measures facilitate mechanisms to "vindicate the norms of due process" and administrative

103. PASQUALE, *supra* note 18, at 14–15; Giovanni Buttarelli, *Towards A New Digital Ethics: Data, Dignity, And Technology*, EUR. DATA PROTECTION SUPERVISOR 2 (Sept. 11, 2015); Rob Kitchin, *Thinking Critically About and Researching Algorithms*, 20 INFO. COMM. & SOC'Y 14 (2017); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. U. L. REV. 1, 21 (2014); Nicholas Thompson et al., *Emmanuel Macron Talks to WIRED About France's AI Strategy*, WIRED (Mar. 31, 2018, 06:00 AM), https://www.wired.com/story/emmanuel-macron-talks-to-wired-about-frances-ai-strategy/ [https://perma.cc/X7HH-4K6K].

104. PASQUALE, *supra* note 18, at 14.

105. Citron, *supra* note 19, at 1310–12.

106. *See* Citron & Pasquale, *supra* note 103, at 21; *see also* Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 125–28 (2014) (advocating a right to "procedural data due process" to address the harms of predictive systems).

107. Kroll et al., *supra* note 20, at 657 (describing transparency as "[a] native solution to the problem of verifying procedural regularity" and describing its utility and limits); Fink, *supra* note 20, at 1453–56 (explaining limits of transparency due to current state of law and practice with respect to whether algorithms would be considered "records" under the Freedom of Information Act (FOIA) and agency bases for withholding algorithms and source code under FOIA requests); Pasquale, *supra* note 20, at 235–36.

108. Citron, *supra* note 19, at 1311–12.

109. *See* Citron & Pasquale, *supra* note 103, at 20–22 (advocating for transparency requirements for data and calculations and placing scoring systems used in the context of employment, insurance, and health care under licensing and audit requirements); *see also* Crawford & Schultz, *supra* note 106, at 122–23.

110. PASQUALE, *supra* note 18, at 14–15; Diakopoulos, *supra* note 44, at 399–402; Citron & Pasquale, *supra* note 103102, at 21–22.

decision making even when decisions are automated.[111] This allows individuals to plead extenuating circumstances that software cannot anticipate[112] and accords the subjects of automated decisions the right to inspect, correct, and dispute inaccurate data.[113]

Yet, while critics have debated the limits of these approaches,[114] the debate has focused largely on the use and effectiveness of transparency, whistleblowers, ex post challenges, and oversight. The ex post focus positions accountability after critical design decisions have been made. And while new scholarship has begun to focus on the process of machine learning system design,[115] this literature has not explored the full potential of administrative law to remedy the abdication of government agencies' involvement in design questions, even when they implicate issues that we usually regard as involving traditional substantive policy questions.

Administrative law maps another direction. It suggests that, when the design of machine learning systems embeds policy, policymakers should be required to engage in reasoned decision making. To be meaningful, given the character of the decisions involved in machine learning design, that

111. Citron, *supra* note 19, at 1301.

112. *Id.* at 1304.

113. PASQUALE, *supra* note 18, at 145.

114. Kroll et al., *supra* note 20, at 657–58 (explaining that while "full or partial transparency can be a helpful tool for governance in many cases . . . transparency alone is not sufficient to provide accountability in all cases"); *see generally* Katherine Noyes, *The FTC Is Worried About Algorithmic Transparency, and You Should Be Too*, PC WORLD (Apr. 9, 2015, 08:36 AM), https://www.pcworld.com/article/2908372/the-ftc-is-worried-about-algorithmic-transparency-and-you-should-be-too.html [https://perma.cc/7KHT-GHZ7]; Christian Sandvig et al., *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, UNIV. MICH. (May 22, 2014), http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf [https://perma.cc/TJ3Y-2UZK] (presenting at "Data and Discrimination: Converting Critical Concerns into Productive Inquiry," a preconference at the 64th Annual Meeting of the International Communication Association). Critiques include the fact that open sourcing a given machine learning system's neural network does not necessarily mean an outside third party will verify how the system determined a given output. *See* Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability*, NEW MEDIA & SOC'Y 973, 983–84 (2016); Jakko Kemper & Daan Koklman, *Transparent to Whom? No Algorithmic Accountability Without a Critical Audience*, INFO. COMM. & SOC'Y (2018); Brauneis & Goodman, *supra* note 6, at 137–38 (pointing out the difficulty of understanding complex AI systems and the shortcomings of knowing inputs and outputs of a given system as the basis for adequate oversight); Maayan Perel & Niva Elkin-Koren, *Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement*, 69 FLA. L. REV. 181, 194–96 (2016). For the impediment posed to transparency by trade secret law, see Brauneis & Goodman, *supra* note 6, at 153–57; David S. Levine, *Secrecy and Unaccountability: Trade Secrets in Our Public Infrastructure*, 59 FLA. L. REV. 135, 180 (2007).

115. *See supra* note 22; *see also* Katyal, *supra* note 20, at 54.

deliberation must address an understanding, informed by both technical and domain expertise, of the methodologies adopted and the value choices behind them, and provide justifications for those choices' resolution. Administrative law, moreover, provides guidance about what types of concerns should trigger such requirements, and how, given the characteristics of machine learning, those concerns translate to the particular context of system design.

B. A FRAMEWORK FOR REASONED DECISION MAKING ABOUT MACHINE LEARNING DESIGN

The administrative state's legitimacy is premised on the foundational principle that decisions of substance must not be arbitrary or capricious.[116] Rather, those decisions must be the product of a contemporaneous process of reasoned decision making.[117] Requiring such process vindicates core public law values: it ensures, on the one hand, that technical expertise has been brought to bear on a decision; and on the other, that the decisional visibility necessary to permit public accountability exists.[118] Together, a transparent reasoning process prohibits an agency from "simply asserting its preference."[119]

Specifically, an agency must produce a record that enables courts "to see what major issues of policy were ventilated," and "why the agency reacted to them as it did."[120] Thus the agency must have engaged in reasoned analysis about relevant factors consistent with the record before it, and they may not have considered irrelevant factors or decided without sufficient evidence. An agency falls short where there is no record of "examin[ing] the relevant data" or "articulat[ing] a satisfactory explanation for its action including a 'rational connection between the facts found and the choice made.' "[121]

By the terms of this standard, the complete abdication of any agency role in considering the important policy choices inherent in a machine learning system's design would be an abject failure. This Section explores the alternative, using the arbitrary and capricious paradigm to identify the types

116. Courts may "hold unlawful and set aside [an] agency action" they deem to be "arbitrary [or] capricious." 5 U.S.C. § 706(2)(A).

117. SEC v. Chenery Corp. (Chenery II), 332 U.S. 194, 196 (1947) (holding that courts may uphold an agency's action only for reasons on which the agency relied when it acted); *see generally* Kevin M. Stack, *The Constitutional Foundations of Chenery*, 116 YALE L.J. 952 (2007) (grounding the *Chenery* norm in the Constitution).

118. Cass R. Sunstein, *From Technocrat to Democrat*, 128 HARV. L. REV. 488 (2014) (discussing the technocratic and democratic directions in administrative law).

119. *Id.* at 496.

120. Auto. Parts & Accessories Ass'n v. Boyd, 407 F.2d 330, 338 (D.C. Cir. 1968).

121. Motor Vehicle Mfrs. Assn. v. State Farm Mut., 463 U.S. 29, 43 (1983) (quoting Burlington Truck Lines, Inc. v. United States, 371 U.S. 156, 168 (1962)).

of machine learning systems, and system elements, whose design should be guided by reasoned and transparent decision making, and what such decision making would require in the machine learning context to survive legal challenge.

1. Determining What System Choices Should Require Reasoned Decision Making

Government agencies increasingly rely on artificial intelligence across their operations. Many functions—from monitoring IT system security to managing government supply lines and procurement—involve largely management support, and therefore may not implicate the types of policy decisions that should trigger the type of decisional record discussed above. This raises a threshold challenge in distinguishing systems that are inward-facing from those that create public-facing policy of the type that agencies should deliberate about and ventilate in a public manner.

Administrative law has dealt with comparable distinctions in a range of contexts and offers some insights into where and how we might draw lines about when a machine learning system is engaged in policymaking of concern to us, and when it is not. Specifically, jurisprudence has identified important indicia of contexts in which administrative choices trigger concerns necessitating a reasoned and transparent decision making process, and the creation of a record sufficient for judicial review: whether the agency action in question limits future agency discretion in deciding issues of legal consequence, and whether the action reflects a normative choice about implementation. Each of these inquiries offer useful insight for the question of which machine learning systems' design, and which system elements' design, should be treated as making policy.

a) Design Choices that Limit Future Agency Discretion

In a variety of contexts, courts have identified the constraining effect of an administrative decision on the future substantive discretion of agencies or their staff as a baseline determinant of whether agency decisions will be subject to judicial review, and therefore to analysis under the arbitrary and capricious standard. When current decisions hem in choices about the law's application going forward, they reflect binding policy choices and thus may be reached openly, explicitly, and through reasoned analysis.

Even in contexts in which executive discretion is broad—as it is in internal agency management—such factors argue for requiring reasoned decision making. Thus, while agencies have largely unreviewable discretion regarding enforcement decisions,[122] judicial oversight is appropriate when an

122. *See* Heckler v. Chaney, 470 U.S. 821, 832 (1985). The court opinion stated:

agency adopts a "general enforcement policy" that "delineat[es] the boundary between enforcement and non-enforcement."[123] Such actions limit agency discretion going forward, with implications for "a broad class of parties."[124] In such contexts, in contrast to individual decisions to forgo enforcement, an agency is expected to present a clearer and more easily reviewable statement of its reasons for acting.[125]

Related concerns govern the determination of whether an agency action is "final," which is a second Administrative Procedure Act (APA) prerequisite for judicial review.[126] To satisfy this requirement, an agency action must not simply mark the "consummation" of an agency's "decision-making process"—a standard satisfied by many nonbinding or advisory decisions, even when they are made informally.[127] The decision must also "be one by which rights or obligations have been determined, or from which legal consequences will flow."[128] The Supreme Court has recently counseled a "pragmatic" approach to the interpretation of this standard, focusing on the prospective limits it places on agency discretion as a key component of the "legal consequences" test.[129] Lower courts have already taken such a pragmatic approach—looking at whether, as a practical matter, a purportedly non-binding agency decision effectively guides future agency decisions and constrains agency discretion, such as if "an agency act[s] 'as if a document issued at headquarters is controlling in the field.' "[130]

The distinctions drawn with respect to finality track those governing whether agency actions must satisfy the notice and comment procedures prescribed by § 553 of the APA. While reasoned decision making sufficient

---

> [W]e recognize that an agency's refusal to institute proceedings shares to some extent the characteristics of the decision of a prosecutor in the Executive Branch not to indict—a decision which has long been regarded as the special province of the Executive Branch, inasmuch as it is the Executive who is charged by the Constitution to "take Care that the Laws be faithfully executed."

Id.; APA excludes from review "agency action . . . committed to agency discretion by law." 5 U.S.C. § 701; *see also* Citizens to Preserve Overton Park, Inc. v. Volpe, 401 U.S. 402, 410 (1971) (holding that the "committed to agency discretion" exception to judicial review is "very narrow" and "is applicable in those rare instances where 'statutes are drawn in such broad terms that in a given case there is no law to apply' ").

123. Crowley Caribbean Trans. v. Pena, 37 F.3d 671, 676–77 (D.C. Cir. 1994).

124. *Id.*

125. *Id.*

126. The APA extends judicial review only to "final agency action." 5 U.S.C. § 704.

127. Bennett v. Spear, 520 U.S. 154, 177–78 (1997).

128. *Id.*

129. *U.S. Army Corps of Eng'rs v. Hawkes Co.*, 136 S. Ct. 1807, 1814–15 (2016); *see* William Funk, *Final Agency Action After Hawkes*, 11 N.Y.U. J.L. & LIBERTY 285 (2017).

130. Appalachian Power Co. v. EPA, 208 F.3d 1015, 1021 (D.C. Cir. 2000).

for system design and adoption decisions to survive arbitrary and capricious review can certainly occur through a range of administrative processes beyond informal rulemaking, this jurisprudence offers an informative framework in which courts have thought carefully about which agency actions should trigger more robust process, reflecting reasoned deliberation, participation, expertise, and judicial review.

In this context, courts have developed extensive doctrine regarding what types of agency actions are "non-legislative" and therefore exempt from such process requirements, as compared to those that are "substantive" and therefore must satisfy them. Such exempt actions (involving, for example, internal agency procedure, agency management, or guidance to regulated parties) do not carry the "force of law" in that they do not make substantive changes to the legal rights and obligations of regulated individuals. As understood by case law, agency guidance statements are those "issued by an agency to advise the public prospectively of the manner in which the agency proposes to exercise a discretionary power."[131] These statements provide agencies with the opportunity to announce their "tentative intentions for the future" in a non-binding manner. An agency articulation, then, that "genuinely leaves the agency and its decision makers free to exercise discretion" raises few process concerns.[132] The agency may adopt it with little process, and it is not, in and of itself, reviewable by courts.

By contrast, courts are also sensitive to the concern that agencies are circumventing the need for decision-making process when they make substantive policy in a manner purported to govern only internal agency procedure or provide only informal guidance. As a result, courts sometimes find that notice-and-comment is necessary, even when the agency statement in question does not seem in and of itself to have any binding legal effect on regulated entities. This seems especially so when the relevant statutes and legislative rules give the agency wide discretion, but the challenged agency statement indicates that agency personnel will in reality exercise that discretion only in narrowly defined circumstances.[133] In those situations, courts have found that the agency action is "practically" (although not formally) binding. Because of the severe constraints that the agency's "informal" action imposed on agency discretion, the agency should have engaged in the full notice-and-comment rulemaking procedure.

131. Am. Bus. Ass'n v. United States, 627 F.2d 525, 529 (D.C. Cir. 1980) (internal citation omitted).

132. *Id.*

133. Gen. Elec. v. EPA, 360 F.3d 188 (D.C. Cir. 2004); Cmty. Nutrition Inst. v. Young, 818 F.2d 943 (D.C. Cir. 1987).

Tracking these standards, existing jurisprudence regarding the setting of formulae and numerical cutoffs, and the choices regarding underlying methodology, provides useful guidance for identifying aspects of machine-learning systems that set discretion-constraining policy.

Pickus v. United States Board of Parole,[134] a case arising in the challenge to an agency's decision to adopt a formula informally (without a notice and comment process), describes well the ways in which the such adoption can set future policy by limiting agency discretion going forward. In *Pickus*, the D.C. Circuit considered a challenge to two rounds of Parole Board "guidelines" that set formulae by which parole would be determined. The court rejected the Board's contention that, under the APA, the issuance of such guidelines lacked legal force because they were merely "general statements of policy, interpretative rules," or "rules relating to agency organization, practice or procedure."[135]

In so doing, the court focused on the practical implications on agency decision-making discretion, and the subsequent legal consequences. As the court described, the first set of guidelines "consist of nine general categories of factors, broken down into a total of thirty-two sub-categories, often fairly specific." Therefore,

> [a]lthough they provide no formula for parole determination, they cannot help but focus the decisionmaker's attention on the Board-approved criteria. They thus narrow his field of vision, minimizing the influence of other factors and encouraging decisive reliance upon factors whose significance might have been differently articulated had [more formal decision-making processes] been followed.[136]

Because of this narrowing of decision-making focus, the court held, the guidelines "were of a kind calculated to have a substantial effect on ultimate parole decisions."

The second agency action, styled an "announcement," consisted of a "complex, detailed table which purport[ed] to state the range of months

134. Pickus v. United States Board of Parole, 507 F.2d 1107 (D.C. Cir. 1974). In a later case, Prows v. United States Dep't of Justice, 704 F. Supp. 272 (D.C. Cir. 1988), a Program Statement from the Federal Bureau of Prisons declaring that inmates had to deposit at least 50% of their payment from prison jobs to "legitimate financial obligations" was struck down. Analogizing the rule to the guidelines in *Pickus*, the court found the Statement "has been interpreted by defendants in a 'formula like' manner," without any discretion and therefore wasn't an interpretative rule nor a policy statement and should have proceeded through notice and comment. *Prows*, 704 F. Supp. at 277.

135. *Pickus*, 507 F.2d at 1112 (D.C. Cir. 1974) (citing 5 U.S.C. § 553(a)(2) and providing exemptions).

136. *Id.* at 1111–13.

which the Board [would] require an inmate to serve depending upon the severity of his offense (six classifications) and his 'salient factor score' (four classifications)."[137] The score, the court continued,

> is computed using only those criteria, and the quantitative input of each is specified as well. Computation of the score is a purely mechanical operation. Third, the chart sets a narrow range of months of imprisonment that will be required for a given category of offense and a given salient factor score. This is not to suggest that these determinants are either unfair or undesirable, but merely that they have significant consequences.[138]

Thus, the court concluded, both policies defining parole selection criteria "are substantive agency action," and "the interested public should have an opportunity to participate, and the agency should be fully informed, before rules having such substantial impact are promulgated."[139]

Moreover, in *Community Nutrition Institute v. Young*,[140] the D.C. Circuit determined that FDA "action levels"—the allowable levels of unavoidable contaminants in food, and again a precise number—while purportedly without the "force of law," practically bound third parties and should have gone through the notice-and-comment procedure required for legislative rules. Pursuant to its statutory mandate to limit the amount of "poisonous or deleterious substances" in food,[141] the FDA established "action levels"—which the FDA characterized as guidance statements—that set permissible levels of unavoidable contaminants such as aflatoxins in food. Producers who exceed action levels are subject to enforcement proceedings. The FDA claimed that action levels were "nonbinding statements of agency enforcement policy," but the court found that setting precise numerical limits cabined the FDA's enforcement discretion, effectively binding the FDA and therefore affecting the rights of regulated parties.[142]

b) Normative Choices Between "Methods of Implementation"

Judge Richard Posner, writing for the Seventh Circuit in *Hoctor v. U.S. Department of Agricriculture*,[143] has articulated the way that numerically-based

137. *Id.* at 1110–11.
138. *Id.* at 1113.
139. Id.
140. Cmty. Nutrition Inst. v. Young, 818 F.2d 943 (D.C. Cir. 1987).
141. 21 U.S.C. § 346.
142. *Cmty. Nutrition Inst.*, 818 F.2d at 946–48 ("[T]his type of cabining of an agency's prosecutorial discretion can in fact rise to the level of a substantive, legislative rule."). That is exactly what has happened here.
143. Hoctor v. U.S. Dep't of Agric., 82 F.3d 165, 171 (7th Cir. 1996).

line-drawing can often reflect a particularly unconstrained form of normative policymaking—which, when is does, enhances the need for more robust process. *Hoctor* involved a challenge to an informal USDA internal memorandum fixing a specific requirement for the height of perimeter fences used to contain "dangerous animals." While the background regulation in force for a number of years had required fencing "appropriate" for the animals involved, the memorandum sought uniformly to require eight-foot fences. The court, however rejected the agency's attempt to arrive at a numerical standard with little decisionmaking process, which, in the court's mind, undermined the decision's democratic legitimacy.[144]

Generally, Judge Posner emphasizes the policy-making nature of administrative decisions that "translate[ ] a general norm into a number"[145]—a phenomenon, he notes, that arises "especially in scientific and other technical areas, where quantitative criteria are common."[146] Moreover, he describes, the "flatter" (or more specific) the ultimate line drawn by the agency, "the harder it is to conceive of it as merely spelling out what is in some sense latent in a statute or regulation"[147] and the more it represents a choice among "methods of implementation."[148] Such choices are legislative in nature, and should be treated as such.

Jurisprudence reviewing agency decision making under the arbitrary and capricious standard reflects these insights about numerical or formula-based agency implementation choices, and provides important foundations for identifying which elements of machine learning systems must satisfy the arbitrary-and-capricious metric in their adoption. Indeed, courts have explicitly held that agencies must engage in reasoned analysis in choosing methods of implementation reflecting many of the very type of decisions inherent in machine learning design described in Part II.

In assessing risk, courts have held, agency decision makers must actively consider the decision whether to err in the direction of false negatives or false positives, and provide reasons for their choice.[149] Similarly, agencies must justify the assumptions behind their use of specific models when

144. SUSAN ROSE-ACKERMAN, STEFANIE EGIDY & JAMES FOWKES, Due Process of Lawmaking: The United States, South Africa, Germany and the European Union 91 (2015).

145. *Hoctor*, 82 F.3d at 171.

146. *Id.*

147. *Id.*

148. *Id.* at 170.

149. *See* Int'l Union, United Mine Workers of Am. v. Fed. Mine Safety & Health Admin., 920 F.2d 960, 962–66 (D.C. Cir. 1990) (remanding to the agency for "more reasoned decision making" on the issue of whether carbon monoxide monitors provide enough protection for workers, after it engaged only in an analysis of false negatives, without discussion of false positives, "ignor[ing] this problem altogether").

determining costs and designing impact statements[150] and provide information to justify the methodology behind models that they use for risk prediction.[151] They must take steps to confirm the validity of their chosen models[152] and, in deciding whether to use a particular scientific methodology, both demonstrate its reliability and transparently discuss its shortcomings. With respect to data, agencies must provide information on its source.[153]

What reasoned deliberation entails is set out across a range of procedural contexts—from cost-benefit analysis to environmental impact assessments—and in a range of substantive policy contexts. The failure to identify, disclose, engage with, and justify the consequent policy choices within models closely correlated to machine learning systems—statistical and economic models, for example—has been held to constitute a "complete lack of explanation for an important step in the agency's analysis."[154] And absent efforts to surface and affirmatively explain the assumptions underlying decision-making models, they may remain "fatally unexplained" and unappreciated.[155]

c) Application to Machine Learning Systems

Decisions about the design of a machine learning system—particularly one modeling fairness—constrain agency discretion much like the formulae in *Pickus*, and the action levels in *CNI*. These cases underscore the ways in which precise numerical limits or formulae have anchoring effects that constrain agency action, and the consequent importance of robust process in their adoption. Machine learning systems are rife with similar issues, such as cutoffs determining who is high, medium, or low risk in recidivism risk systems, or the thresholds in the Amazon Rekognition service described in Part II.

150. Nat. Res. Def. Council, Inc. v. Herrington, 768 F.2d 1355, 1412–19 (D.C. Cir. 1985) (analyzing the Department of Energy's use of a real annual discount rate of 10% when determining life cycle costs and the net present value of savings from appliance energy efficiency standards).

151. Owner-Operator Indep. Drivers Ass'n, Inc. v. Fed. Motor Carrier Safety Admin., 494 F.3d 188, 199–204 (D.C. Cir. 2007) (holding that an agency must disclose the methodology of the agency's operator-fatigue model, a crash-risk analysis that was a central component of the justification for the final rule).

152. Ecology Ctr. v. Austin, 430 F.3d 1057 (9th Cir. 2005). The Forest Service used a model to conclude that treating old-growth forest through salvage logging was beneficial to dependent species but did not confirm their hypothesis by any on-the-ground analysis.

153. *Nat. Res. Def. Council*, 768 F.2d at 1412–19.

154. *Owner-Operator Indep. Drivers Ass'n, Inc.*, 494 F.3d at 204.

155. *Natural Res. Def. Council*, 768 F.2d at 1414–19 (analyzing the Department of Energy's use of a real annual discount rate of 10% when determining life cycle costs and the net present value of savings from appliance energy efficiency standards).

There is, moreover, often no unique connection between the cutoffs or thresholds chosen and a statutory mandate or technological requirement, as in *Hocter*. Rather—like the agency actions in the arbitrary-and-capricious cases involving which risk models to adopt, whether to prefer false negatives and positives, what data to use, and which scientific methodology to employ—those decisions reflect normative choices between methods of implementation. In the machine learning context, these might also include the unit of analysis (the algorithm, the algorithmic system, or the overall system of justice);[156] how fairness is measured—whether it is by group-level demographic parity, equal positive predictive values, equal negative predictive values, accuracy equity, individual fairness metrics such as equal thresholds, or a devised similarity metric—and the related question of whether and how to use attributes related to protected classes such as in the *Loomis* case.

When the design of a machine learning system deprives an agency and its staff of future substantive discretion,[157] especially through numerical or methodological choices that reflect normative judgments on implementation rather than ones required directly by statute or technical or scientific knowledge, the design choices embedded in machine learning systems should not be reached in an arbitrary or capricious manner. Thus if a record lacks evidence of agency deliberation or reveals deliberations that demonstrate one of the other indicia of arbitrariness, an agency's reliance on the system should be subject to legal challenge.

2. *Designing Agency Decision Making: Reflecting the Technocratic and Democratic Requirements of Administrative Law*

Where the policy decisions embedded in system design supplant administrative discretion, what would it mean, in the language of the arbitrary and capricious jurisprudence, that on the one hand "the agency should be fully informed"[158] and provide a justification for its choice based on a "rational connection" with the "facts found,"[159] and on the other, that decisions should be open to public engagement and political accountability?

156. Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, *in* PROC. CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 59, 60–61 (2019) (describing the "framing trap"—the tendency to analyze fairness at the level of inputs and outputs of the model rather than at the level of socio-technical system in which the machine learning system is embedded).

157. Am. Bus. Ass'n v. United States, 627 F.2d 525, 529 (D.C. Cir. 1980) (asking "whether a purported policy statement genuinely leaves the agency and its decision-makers free to exercise discretion").

158. Pickus v. United States Board of Parole, 507 F.2d 1107, 1113 (D.C. Cir. 1974).

159. Motor Vehicle Mfrs. Assn. v. State Farm Mut., 463 U.S. 29, 43 (1983) (quoting Burlington Truck Lines, Inc. v. United States, 371 U.S. 156, 168 (1962)).

These elements reflect dual (and sometimes competing) impulses—technocratic and democratic—animating the law of administrative process.[160]

As to the first, to engage in reasoned deliberation, agency staff must address their lack of technical knowledge, enlist additional expertise to "inform" themselves sufficiently, and provide reasons justifying the resolution of four questions specific to machine learning systems. Those questions include: (1) for what a system is optimizing; (2) what determinations are being made about the choice and treatment of data; (3) what assumptions and limitations are implied by the choice of model; and (4) what interfaces and policies structure agency staff's interactions with machine learning systems—the human-machine loop. Importantly, as to the second question, meaningful processes must address the opacity of value choices made through design by ensuring "political visibility,"[161] to surface the fact that technical choices involve a policy judgment. In this context, transparent decision making involves not simply making algorithms transparent, but making policy visible.[162]

a) Technocratic Elements in Reasoned Decision Making About Machine Learning Systems

A comparison of machine learning systems with prior automation—generally so-called "expert systems"—helps identify particular aspects of machine learning system design decisions that displace traditional modes of expert administrative decision making. Danielle Citron's 2008 work on *Technological Due Process*[163] provided a foundational analysis of the ways that administrative automation based on a prior generation of expert systems transformed the technological decision-making landscape in ways that matter for policymaking norms.[164] It is further instructive in highlighting the ways machine learning has both compounded and redirected the displacement of expert human judgment, a challenge with which agencies must grapple when adopting such systems.

160. *See generally* Sunstein, *supra* note 118 (discussing the technocratic and democratic strains in administrative law).

161. Mulligan & Bamberger, *supra* note 12, at 776–80.

162. *See, e.g.*, Eaglin, *supra* note 22, at 88 (noting how recidivism risk tools make it "difficult to ascertain . . . policy decisions").

163. Citron, *supra* note 19.

164. For an excellent and accessible discussion of expert systems and what lessons from their development suggest about the discussion for explainability and interpretabiliy in machine learning, see generally David C. Brock, *Learning from Artificial Intelligence's Previous Awakenings: The History of Expert Systems*, 39 AI MAG. 3 (2018).

i) Citron's Concerns: Displacement of Expert Agency Judgment

Citron identified a related set of objections to earlier attempts to automate agency processes. First, she described how "[a]utomated systems inherently apply rules because software predetermines an outcome for a set of facts."[165] This, in turn, displaces the ongoing exercise of human judgment, which is better reflected in standards. She thus concludes that "[d]ecisions best addressed with standards should not be automated."[166] Citron further drew on the "rules versus standards" debate to emphasize the distinction between automated systems, which implement rules and favor consistency, and human decision-making systems, which favor "situation-specific discretion."[167]

Second, Citron raised the related question of *who* sets the rules that displace the standards-like exercise of human judgment. Her concern involved the displacement of expert agency decision making by the choices of engineers who design technical systems.[168] In particular, she was concerned that engineers' interpretations and biases, and their general preference for tractable binary questions, distort decision making.

Finally, Citron expressed concern regarding the lack of record-keeping and transparency about the rules automated systems apply. Absent such a digital trail, the ability to seek redress or accountability is limited. To enable individual due process and support overall accountability, Citron advocates that systems be built to produce "audit trails that record the facts and rules supporting their decisions."[169]

ii) Updating Concerns: How Machine Learning Displaces Rational Expert Agency Decision Making

While Citron's conception of what is inherent in automation may have been largely accurate with respect to the automated systems used by government at the time (prior to her 2008 publication date), the rote

165. Citron, *supra* note 19, at 1303.

166. *Id.* at 1304.

167. *Id.* at 1303; *see* Bamberger, *supra* note 18, at 676 ("Computer code [in contrast to human judgment] operates by means of on-off rules, while the analytics it employs seek to quantify the immeasurable with great precision.") (internal quotation marks omitted).

168. Citron, *supra* note 19, at 1261 ("Code writers also interpret policy when they translate it from human language to computer code. Distortions in policy have been attributed to the fact that programmers lack 'policy knowledge.' "); *id.* at 1262 ("Changes in policy made during its translation into code may also stem from the bias of the programmer. . . . Policy could be distorted by a code writer's preference for binary questions, which can be easily translated into code.").

169. *Id.* at 1305.

application of predetermined rules she documents is an inapt description of the machine learning systems coming into government use today. Machine learning systems do not apply predetermined rules to sets of facts, but rather develop probabilistic models that optimize for a particular goal. They are then allowed to learn in the field, generate new rules on the fly, and iteratively update them.

In this way, like earlier expert systems, machine learning systems too displace agency reasoning and expertise, and constrain future agency discretion. However, the displacement takes new forms, stems from additional sources, and requires distinct responses. The risk of displacement no longer stems from the explicit reasoning of engineers translating agency rules into code, but rather arises from the "logic" the model machine learning systems derive from training data reflecting past agency actions. The assumptions and policy choices built into the machine learning model used to generate the predictive model, as well as policy choices in the application of the predictive model, rather than engineer-coded rules, are the key hidden constraints.

a. Element 1: Delegating "Logic-Making" to Machines

Today's machine learning systems, then, delegate "logic-making" to algorithms. Unlike expert systems that Citron rightly identified as displacing nuanced and fact-specific agency staff decisions with the rote application of predetermined rules as coded by engineers, machine learning systems *construct their own logics* from training data. Machine learning systems skip the process of codifying an agency's decision-making process, and instead rely on the machine learning model to learn a classifier—its own machine logic—from a set of training data that reflects past agency actions. While the machine logic captured in the classifier could be considered more analogous to the intuition and instinct associated with agency experts,[170] importantly, it in no way reflects the logic of agency decision makers. In fact, it answers without the causal reasoning associated with logic, or as one scholar notes, "they don't 'think' in any colloquial sense of the word—they only answer."[171]

While many consequential decisions are made by the engineers, the decision about how to model agency judgments is not explicitly constructed by engineers through rules—abstract or specific—but rather learned by

170. Of course, human intuition is produced by neurological processes and machines' through computational processes. While machine learning abounds with terms that evoke the brain, only some machine learning systems attempt to mirror cognitive processes.

171. Jonathan Zittrain, *The Hidden Costs of Automated Thinking*, NEW YORKER (July 23, 2019), https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking [https://perma.cc/D9YK-JYHZ].

algorithms through analysis of data traces reflecting agency decision making. In theory, one might surmise that because machine learning models are trained on data that represents the past decisions and related outcomes of the agency—or "similar" ones—they might naturally align more closely with the judgment of agency experts and, by design, provide less room for interference or usurpation of judgment by engineers. If that were so, perhaps machine learning systems should raise *less* concern about displacement of human expert judgment than earlier automated systems.

Unfortunately, this surmise breaks down under more careful scrutiny. The training data reflects patterns reflected in professional decisions, but not professionals' *decision-making processes*. This is an important distinction. It means that a machine learning model's "logic" may well reflect the actions and outcomes of professional decision making (the outputs) but bear little resemblance to the rationales and justifications behind those decisions.

Significantly, the "reasoning" of complex machine learning systems often bears no resemblance to human logic and is impossible to discern.[172] The divergence in intuition is intrinsic because machines and humans "see" in different ways. For example, machine learning systems can identify complex patterns and scan across massive data sets. Humans, by contrast, can identify things they've seen (such as faces) despite a wide range of subtle and relatively extreme perturbations (changes to hair style, plastic surgery, aging, etc.). The different intuitions developed by human and machine systems may therefore produce similar outputs in some instances, but not in others, or similar outputs but for very different reasons.

Thus, the machine has learned its own logic based on the training data: it has not learned to mirror the agency's logic, only to predict outcomes of it. Like two students producing correct answers to a math problem, unless they "show their work," it would be wrong to assume they used the same method, let alone that either used the right method or appropriately applied it. The design process of machine learning systems does not explicitly transfer expert reasoning and therefore does not create the pattern of displacement found in expert systems. Yet because machine learning classifiers are developed by studying the outcomes of agency logics rather than the logic itself, it creates potentially more troubling displacement effects.

172. For example, machine learning image recognition systems are famous for appearing to perform well on a task but actually relying on a simplistic and poorly-chosen heuristic. In addition, as Kaminsky describes, algorithms lack the contextual understandings of acceptable bases for decision making and the common sense of humans. Kaminsky, *supra* note 22, at 14–15.

b. Element 2: Constraints on Policymaking Evolution

Machine learning systems develop *probabilistic models that optimize for a particular goal*—and then, where they are allowed, update them as they learn in the field. Rather than the automated rules that concerned Citron, the constraints imposed on agency discretion in machine learning systems are found in choices about what a system is optimizing for and how the goal is operationalized going forward within the system.

Once deployed, the logic of the model—whether fixed, or allowed to learn over time—remains constrained by the assumptions and choices made during design. In contrast, the judgment of agency professionals and staff may evolve over time, sometimes on a gentle slope, but at other times diverging swiftly in response to new research, new political winds, or other internal or external jolts. While a machine learning model may learn new ways to optimize for the goal established, it is tethered to the beliefs and biases that are fixed in the model, as well as the assumptions and ingested data used during development. As a result, machine learning systems can instill patterns of racism, debunked science, or other faulty or unjust reasoning that may be captured in the training data or optimization choices.

Even if the policies embedded in a model are fully aligned with agency decision making at the time of its initial deployment, if the system is not updated to reflect changing agency understanding of sound judgment and agency practice, machine learning systems can constrain agency discretion in particularly problematic ways.

Finally, the extent to which a model developed on one data set can be safely used on another is an immensely important policy question. It is well documented that models trained on one data set can perform catastrophically poorly on a data set that many might assume to be similar by some set of metrics.[173] For example, models trained on newswire copy perform poorly on texts from other domains.[174] Even for discrete Natural Language Processing (NLP) tasks, such as identifying words as nouns, verbs, adjectives, etc.—called part-of-speech or grammatical tagging or word-category disambiguation—which lay people might consider simple and transferable

173. Selbst et al., *supra* note 156, at 4–5 (calling this the "portability trap" and tying it to the quest for abstractions and tools that can be reused across contexts).

174. *See* David Bamman, *Natural Language Processing for the Long Tail*, DIGITAL HUMAN. 2 (2017) ("[T]he performance of an out-of-the-box part-of-speech tagger can, at worse, be half that of its performance on contemporary newswire. On average, differences in style amount to a drop in performance of approximately 10–20 absolute percentage points across tasks.").

across corpora, models trained on news articles perform quite poorly on literary works.[175]

iii) The Challenge: Reintroducing Expert Justification for Agency Decisions

The way in which machine learning systems generate decisions without decision-making processes challenges administrative law's fundamental mandate of reasoning. To be legitimate, reliance on machine learning in governance requires processes that reintroduce appropriate expertise in providing justifications for administrative choices.

The requirement of "justification" regarding a system's design and its subsequent choices is critical. Justification is distinct from two elements identified by computer scientists related to system accountability: interpretability—properties or qualities or techniques related to a system that help humans understand the relationship between inputs and outputs[176]—and explainability: the ability to explain the operation of a machine learning system in human terms.[177] Explainability provides the reasoning behind the relationship between inputs and outputs interpretability reveals.[178]

While both interpretability and explainability might be helpful, they are not sufficient to satisfy administrative legitimacy.[179] Explaining an algorithm's operation without providing informed justifications for the choices reflected in that operation fails the "arbitrary and capricious" threshold. Instead,

175. *See id.* (summarizing research investigating the disparity between training data and test data for several NLP tasks).

176. Finale Doshi-Velez & Been Kim, *Towards a Rigorous Science of Interpretable Machine Learning*, ARXIV (Mar. 2, 2017), https://arxiv.org/pdf/1702.08608.pdf [https://perma.cc/6CVL-DWRK].

177. Explanations can describe the operation of a model in general (so-called "global" explanations) or for a particular mechanism in the model used to relate inputs and outputs (so-called "local" explanations). Upol Ehsan et al., *Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations*, ARXIV (Dec. 19, 2017), https://arxiv.org/pdf/1702.07826.pdf [https://perma.cc/GK5X-MXC5].

178. Both explainability and interpretability are areas of debate and research among computer scientists and the multiple disciplines within the broader "fairness, accountability and transparency" research community. For a discussion of these terms and others within and across relevant disciplines, see Nitin Kohli et al., *Translation Tutorial: A Shared Lexicon for Research and Practice in Human-Centered Software Systems*, CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY (2018).

179. For a discussion of the relationship between explanations and justifications in criminal law, and probable cause in particular, see Kiel Brennan-Marquez, *"Plausible Cause": Explanatory Standards in the Age of Powerful Machines*, 70 VAND. L. REV. 1249, 1288 (2017) ("Apart from safeguarding constitutional values, explanations also vindicate rule-of-law principles. A key tenet of legality, separating lawful authority from ultra vires conduct, is the idea that not all explanations qualify as justifications.").

design choices that embed policy choices must reflect reason, a rational connection to the facts, context, and the factors mandated by Congress in the relevant organic statute, while avoiding elements extraneous to the legislative command. And such justification, in turn, requires the application of a range of forms of expertise, including technical knowledge about machine learning and algorithm design, as well as statistics, domain expertise, and specialized fields such as those represented in the Fairness, Accountability, and Transparency in Machine Learning (FAT*) and ethics, law and sociology communities, whose members investigate the social and political consequences of algorithmic systems.

As an initial matter, when systems are adopted by governments, agencies must be able to enlist sufficient expertise at the design phase to permit knowledgeable exploration of technical design and data choices that embed policy. As discussed above, decisions about system goals (what is optimized for), how to operationalize the goal into a target variable for the system to optimize for, and what modeling frameworks to use, all require expert input because they are fundamental policy decisions. In addition, determinations about the data—its selection, curation, cleaning, and similarity to the data on which it will be used—and about the triggers for updating or replacing it are all essential policy questions with which agencies must grapple explicitly. Decisions about the use and inclusion of data about protected traits warrant particular scrutiny. Precise numerical limits such as cut-offs or thresholds—particularly those that cabin discretion—must be the product of reasoned agency decision making.

Additionally, consistent with the case law's emphasis on agency discretion, agencies must comprehend and address the impact of a system on future agency choices. Traditionally, agency staff are able to adjust to new informational inputs as a situation requires.[180] They can selectively pull data in and out of the decision-making frame based on case-specific, situational knowledge. Machine learning—like other automated systems—can constrain the ability to flexibly alter the data brought to bear on a decision in response to the particular problem or person presented.[181] While machine learning systems can process tens of thousands of data points, they can only consider the data predetermined to be relevant. Setting bounds on what can be considered—ensuring, for example, that information about race, gender, age, or other protected attributes does not infiltrate agency decision making—may align with a simplistic notion of fairness. But using such simple

180. *Cf.* Kaminsky, *supra* note 22, at 13–14 (describing how moving from a human to an automated decision can eliminate "cultural knowledge about what is or is not an appropriate decisional heuristic in a particular case").

181. *See* Citron, *supra* note 19, at 1304 (explaining that policies allowing "individuals to plead extenuating circumstances that software cannot anticipate" should not be automated).

categories has been found to frequently be at odds with justice, the goal it purportedly serves.

Even where systems are billed as "decision support," ostensibly allowing decision makers to consider other information, automation bias may lead to overreliance on machine outputs.[182] Without efforts—policy, system design, and accountability frameworks—to foster questioning, agency staff may come to defer to machine outputs, particularly over time. In doing so, systems may elevate ideals of procedural fairness at the cost of substantively just and right outcomes. Angele Christin's research documents that automation bias may not always result and suggests that this tension between different visions of fairness may be a point of resistance. She found different kinds of resistance and tinkering with recidivism risk tools in the justice system—some of which appears to be grounded in battles over competing conceptions of fairness, its relation to justice, and the role that discretion, rather than rigidity, plays in advancing the latter.[183] The risks posed by automation bias nevertheless loom large when relevant professional, regional, or site-specific experts are not consulted during system development,[184] or when the systems are acquired as commercial off-the-shelf products rather than collaboratively developed or tailored for the conditions and context of use.

Because of this limited input and the ways these systems constrain agency staff's ability to expand or narrow the data used to render a decision, and to shift their reasoning over time, machine learning systems risk upsetting context-specific, domain-specific, and evolving judgments—key rationales for agency existence. For these reasons, the interfaces and policies that structure agency staff's interactions with machine learning systems must be the subject of agency deliberation and involve reasoned application of

182. *See* Kate Goddard et al., *Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators*, 19 J. AM. MED. INFORMATICS ASS'N 121 (2011) (reviewing literature on automation bias in health care clinical decision-support systems).

183. Angèle Christin, *Algorithms in practice: Comparing web journalism and criminal justice*, BIG DATA & SOC'Y 1, 9 (July 16, 2017) (discussing a senior judge's perspective on recidivism risk tools: "I don't look at the numbers. There are things you can't quantify . . . [y]ou can take the same case, with the same defendant, the same criminal record, the same judge, the same attorney, the same prosecutor, and get two different decisions in different courts. Or you can take the same case, with the same defendant, the same judge, etc., at a two-week interval and have completely different decision. Is that justice? I think it is" and finding probation officers similarly resisting rigidity by tinkering with the criteria to obtain the score they thought adequate for a given defendant).

184. For example, criminal justice risk-assessment tools, which have been around for decades and are often simply logistic regressions, are almost uniformly created outside of the jurisdictions in which they are deployed. There are fewer than sixty tools used across the entire United States. Angèle Christin et al., *Courts and predictive algorithms*, DATA & CIV. RTS.: A NEW ERA POLICING & JUST. (Oct. 27, 2015).

expertise about the human-machine loop. This includes agency policies of the type Citron recommends—such as training on automation bias and requiring explanations of the facts and findings produced by automated systems on which agency staff rely[185]—as well as decisions about system interfaces, such as whether to communicate uncertainty and, if so, how to do so.

b) Democratic Elements in Reasoned Decision Making About Machine Learning Systems

In addition to gathering the expertise necessary to understand, explain, and justify these design choices, the arbitrary and capricious jurisprudence points to deeper issues about what meaningful deliberation would require in the machine learning context. Specifically, its emphasis on the public disclosure of the decisions made and the assumptions behind them reflects the reality that "[m]odels and proxies are built on numerous assumptions, often based in scientific principles but also laden with value judgments."[186] As political scientist Shiela Jasanoff describes, "there is growing awareness that science cannot answer all of our questions about risk and that both scientific and value judgments are involved in the processes of risk assessment and risk management."[187]

Agencies cannot create a meaningful record of pertinent "issues of policy" involved in machine learning system design and "why the agency reacted to them as it did"[188]—indeed they cannot be transparent to the public, if they fail to disclose both information about the code and its underlying models, limits, defaults, assumptions, training data, and the very fact that they engaged in a policy judgment and how those judgments were resolved. Decisional transparency must involve not only openness about design but also publicity about the very existence and political nature of value questions being resolved through design processes.

Thus "political visibility,"[189] rather than algorithmic transparency, is the essential characteristic of legitimate processes for adopting complex algorithmic systems. Administrative legitimacy is predicated on the explicit public articulation of value choices under consideration and transparent

185. Citron, *supra* note 19, at 1306–07.

186. Sara A. Clark, *Taking a Hard Look at Agency Science: Can the Courts Ever Succeed*, 36 ECOLOGY L.Q. 317, 331 (2009).

187. Sheila Jasanoff, *Cultural Aspects of Risk Assessment in Britain and the United States*, *in* THE SOCIAL AND CULTURAL CONSTRUCTION OF RISK 359, 359 (B. B. Johnson & V. T. Covello eds., 1987).

188. Auto. Parts & Accessories Ass'n v. Boyd, 407 F.2d 330, 338 (1968).

189. Mulligan & Bamberger, *supra* note 12, at 251.

deliberation about their resolution.[190] When values are embedded in design choices they are "less visible as law, not only because it can be surreptitiously embedded into settings or equipment but also because its enforcement is less public."[191] The regulative features of technology design can appear "constitutive"—non-normative and part of the natural state of things.[192] If they are not explicitly surfaced (as they often are not), the policy decisions built into machine learning systems "fade into the background and hide the political nature of [their] design."[193] Value trade-offs, unrecognized as governance, remain unaddressed at the design stage, hindering both robust consideration of substantive policy and ex post oversight.

IV. BUILDING ADMINISTRATIVE PROCESS FOR MACHINE LEARNING

Reasoned decision making about machine learning system adoption requires both deep subject-matter expertise, a key grounding for delegating the power to implement and enforce laws to agencies,[194] and processes ensuring that policies embedded in system design appear and remain politically salient to agency employees as well as to the public and the political branches. Unconsidered resolution of policy issues—including those impacting protected classes—constitutes the epitome of arbitrary and capricious decision making and an abdication of policymaking responsibility at the heart of administrative legitimacy,[195] which displaces expert agency judgment with algorithmic output. Furthermore, this disappearance of values can unintentionally lead agencies that are heavily dependent on machine learning systems to ossify policies that no longer serve the agency's interests and goals.

---

190. *See, e.g.*, *Boyd*, 407 F.2d at 338 (D.C. Cir. 1968) (noting that an agency rulemaking record must make visible "what major issues of policy were ventilated" and "why the agency reacted to them as it did").

191. Lee Tien, *Architectural Regulation and the Evolution of Social Norms*, 7 YALE J.L. & TECH. 1, 22 (2004).

192. Mireille Hildebrandt, *Legal and Technological Normativity: More (and Less) than Twin Sisters*, 12 TECHNE 169, 179 (2008).

193. Mulligan & Bamberger, *supra* note 12, at 778.

194. Cass R. Sunstein, *Constitutionalism After the New Deal*, 101 HARV. L. REV. 421, 442–44 (1987) (discussing the "New Deal belief in the importance of technical expertise" as a justification for according agencies "a large measure of autonomy").

195. *Cf.* Jody Freeman & Adrian Vermeule, *Massachusetts v EPA: From Politics to Expertise*, SUP. CT. REV. 51 (2007) (discussing the Supreme Court's "expertise-forcing" jurisprudence ensuring that "agencies actually do exercise expert judgment"); Heckler v. Chaney, 470 U.S. 821, 833 n.4 (1985) (applying arbitrary and capricious review, even in enforcement contexts characterized by high executive discretion, when an agency's failure to exercise its discretion "amount[s] to an abdication of its statutory responsibilities").

These pitfalls have particular resonance when policy is driven by machine learning system design, where the metrics by which legal rights and obligations are fixed in individual cases are dynamic, and where forms of localization are necessary for performance, including on values such as fairness. Learning systems "learn." Their analytics and algorithms evolve and change according to the logics that machines induce by observing human actions.[196] The policies that these systems implement will change over time and be driven by machine rather than human reasoning, in a way that displaces the discretion of agency staff going forward.[197]

A key justification for delegating substantive policy choices to agencies, of course, is their ability to revise policy "in light of evolving societal, political, and technological circumstances."[198] Yet when those revisions generate legal effects, administrative law requires engagement, reasoning, and transparency.[199] The challenge of public machine learning adoption, then, is to ensure such process as policy is made on a continuum—at design time, configuration time, and run time.[200]

While a range of agency processes might address the opacity of complex machine learning systems and account for the technocratic and democratic demands of reasoned governance,[201] this Part recommends elements for a framework of public machine learning adoption that satisfies both.

---

196. *See supra* Section III.B.1 (discussing decision making by machine learning systems).

197. *See* Mulligan & Bamberger, *supra* note 12 (discussing further the ways that values in technology change over time as technology is appropriated by users in new and unexpected ways, and how technology interactions with business models, organizational structures, and other technologies in ways that can transform its effects, use, and impact on values); Harry Surden, *Structural Rights in Privacy*, 60 SMU L. REV. 1605 (2007) (discussing ways that technology affects the "latent structural constraints" that work to protect values in addition to and in conjunction with legal measures).

198. Kenneth A. Bamberger, *Provisional Precedent: Protecting Flexibility in Administrative Policymaking*, 77 N.Y.U. L. REV. 1272, 1280 (2002); *see* Matthew C. Stephenson, *Public Regulation of Private Enforcement: The Case for Expanding the Role of Administrative Agencies*, 91 VA. L. REV. 93, 139 (2005) ("Flexibility, like expertise, is often invoked to justify delegation of substantive policy choices to agencies.").

199. *See, e.g.*, FCC v. Fox Television Stations, Inc., 556 U.S. 502 (2009) (applying arbitrary and capricious review to a change of agency policy applied in an adjudication); Motor Vehicle Mfrs. Assn. v. State Farm Mut., 463 U.S. 29, 30 (1983) (applying arbitrary and capricious review to a change in agency policy reached through rulemaking).

200. Clark et al., *supra* note 50, at 463 (discussing how values tussles play out at design, redesign, configuration, and run time). This "developer" perspective on the ability and need to address values at every stage of the process is captured in the security adage, "Secure by Design, Secure by Default, Secure in Deployment." Steve Lipner, *The Trustworthy Computing Security Development Lifecycle*, 20TH ANN. IEEE COMPUTER SECURITY APPLICATIONS CONF. (2004).

201. *See* TAXPAYER ADVOCATE SERV., *supra* note 60.

At its core is the reliance on centers of expertise—on the model of the USDS and the 18F "skunk works" team first developed by the Obama Administration—that develop and provide shared technical knowledge in ways that address expertise gaps across agencies while providing a systemic approach to the use of technology in government activity. We further identify different tools that such a centralized effort should employ, including algorithmic impact assessments, which not only involve deliberation about technical choices themselves, but also surface their policy implications publicly.

This framework uses two critical means to build on that visibility and foster public participation, political oversight, and informed agency engagement, during both system design and deployment. The first is institutional. It suggests that the adoption of processes that engage the public as policy is made through design. The second is technical. It suggests that reasoned agency deliberation about policy requires that machine learning systems adopted by governments reflect "contestable" design from the start—design that supports meaningful contestability throughout the system lifecycle by permitting an ongoing role for agency staff in shaping the policies embedded in systems.

Together, these tools focus on the development of expertise and the surfacing of politics while emphasizing judgment, coherence, efficiency, and transparency in setting administrative policy.

A. INFORMING AGENCY DELIBERATION WITH TECHNICAL EXPERTISE

1. Reviewing Piecemeal Efforts

In other works, we have argued that technology should be used to govern only when an agency has access to relevant technical expertise and the ability to consider a wide scope of public values that may be implicated in the use of technology to regulate.[202] We identified a range of options for acquiring relevant expertise, including agency hiring, drawing on the expertise of other agencies, and soliciting expertise from external stakeholders.[203]

Flavors of these alternative approaches to leveraging expertise are visible in some novel processes around algorithmic systems set up by select agencies and pending legislative proposals that seek to address both the potential for inherent bias and privacy risks.

Several jurisdictions at all levels have chosen to create public or quasi-public bodies to aid in the analysis of algorithmic-system adoption. At the

202. Mulligan & Bamberger, *supra* note 12, at 759, 768–70. We also note that stakeholders must have the technical expertise to meaningfully participate and suggest alternative models for providing it. *Id.* at 775–76.

203. *Id.* at 768–70.

local level, New York City passed an algorithmic accountability bill assigning a task force to examine the way city government agencies use algorithms.[204] New York's Automated Decision Systems Task Force, comprised of a cross-disciplinary group of city officials and outside experts,[205] is tasked with recommending a process for reviewing the city's use of automated decision systems to ensure equity and opportunity. The Task Force has held two public workshops to date.[206] While the law brings experts in to assist the city in developing review processes and conducting reviews, in recent testimony submitted to the New York City Council Committee on Technology, two Task Force members wrote:

> Task Force members have not been given any information about ADSs [automated decision systems) used by the City. To date, the City has not identified even a single system. Task Force members need to know about relevant systems used by the City to provide meaningful recommendations.[207]

They further reported that the Task Force had nevertheless made "meaningful progress in developing a methodology for eliciting relevant information about ADSs, using so-called "ADS Cards" that ask developers and operators to provide specific details about the system in question," but that the City forced them to abandon the project.[208]

At the state level, the Pennsylvania Commission on Sentencing[209] has brought expertise into the creation of their recidivism risk system in numerous ways. The Commission was tasked with creating a recidivism risk tool—initially paper-based but over time governed by software.[210] To

204. *See* Int. No. 1696-A, Automated decision systems used by agencies (NYC 2018), http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0 [https://perma.cc/8AZR-DYB9].

205. *Members*, NYC AUTOMATED DECISION SYSS. TASK FORCE, https://www1.nyc.gov/site/adstaskforce/members/members.page [https://perma.cc/R5VZ-LMQL] (last visited Oct. 10, 2019).

206. *Past: Forum #2: Transparency*, NYC AUTOMATED DECISION SYSS. TASK FORCE, https://www1.nyc.gov/site/adstaskforce/events/events.page [https://perma.cc/LPR2-73JV] (last visited Oct. 10, 2019).

207. *Testimony regarding Update on Local Law 49 of 2018 in Relation to Automated Decision Systems (ADS) Used by Agencies before NYC Council Comm. on Technology* (Apr. 4, 2019) (testimony of Julia Stoyanovich & Solon Barocas), https://dataresponsibly.github.io/documents/StoyanovichBarocas_April4,2019testimony.pdf [https://perma.cc/KY2W-M6PY].

208. *Id.*

209. The Commission was created in 1978 to develop and oversee the development of statewide guidelines to promote fairer and more uniform sentencing.

210. Judicial Code and Prisons and Parole Code, 95 PA. STAT. ANN. §§ 42, 61 (2010) (providing for the adoption of a risk assessment tool).

develop these tools, the Commission has conducted its own research,[211] partnered with academic institutions to hold workshops,[212] commissioned academic research,[213] and most recently commissioned an independent evaluation by the Urban Institute.[214]

At the federal level, the Organ Procurement and Transplantation Network (OPTN) and Task Force on Organ Procurement and Transplantation, which were established to regulate the donation and allocation of organs through the network,[215] have a lengthy history of engaging experts in the design of their organ allocation systems.[216] The

211. For overview of research, see *Research Overview of the Sentence Risk Assessment Instrument*, PA. COMMISSION ON SENT'G 4–5 (Oct. 2018), http://pcs.la.psu.edu/guidelines/proposed-risk-assessment-instrument/additional-information-about-the-proposed-sentence-risk-assessment-instrument/research-overview-of-the-sentence-risk-assessment-instrument-1/view; for reports to date, see PA. COMMISSION ON SENT'G, http://pcs.la.psu.edu/publications-and-research/risk-assessment [https://perma.cc/RH9P-CNRU] (last visited Aug. 1, 2019).

212. *See Pennsylvania Criminal Justice Roundtable*, PENNSTATE CRIM. JUST. RES. CTR. (May 19–20, 2011), https://justicecenter.la.psu.edu/research/projects/pennsylvania-criminal-justice-roundtable [https://perma.cc/WFQ7-JF27] (convening state criminal justice policymakers and experts in offender risk assessment and sentencing, including faculty at the Penn State Criminal Justice Research Center).

213. Matthew DeMichele & Julia Laskorunsky, *Sentencing Risk Assessment: A Follow-up Study of the Occurrence and Timing of Re-Arrest among Serious Offenders in Pennsylvania*, PA. COMMISSION ON SENT'G (May 2014), https://justicecenter.la.psu.edu/research/projects/files/PCS%20_Risk%20Assessment_Tool.pdf/view [https://perma.cc/PZ9T-TB5N] (analyzing the relationship between offender and case characteristics and likelihood of recidivism).

214. *See* GEN. ASSEMB. COMM'N ON SENTENCING, PROPOSED SENTENCE RISK ASSESSMENT INSTRUMENT FOR 204 PA. CODE CHAPTER 305; RESPONSES TO PUBLIC COMMENTS; REQUEST FOR PROPOSALS; 48 Pa.B. 5445, at 2 (Sept. 1, 2018) (reporting that "in April 2018, following publication of a revised proposal, staff provided the Urban Institute with a complete set of files related to construction of the instrument (e.g., data, syntax, etc.) to begin the external review").

215. National Organ Transplant Act, Pub. L. No. 98-507, 98 Stat. 2339 (1984) (establishing the Organ Procurement and Transplantation Network and banning organ sales). For a detailed legislative history covering the process and substantive considerations leading up to enactment, see Jed Adam Gross, *E. Pluribus UNOS: The National Organ Transplant Act and Its Postoperative Complications*, 8 YALE J. HEALTH POL'Y L. & ETHICS 145, 207–22 (2008).

216. *See* Gross, *supra* note 215, at 228–30 (describing the NOTA Task Force on Organ Transplantation, which included "medical professionals, social and behavioral scientists, a legal scholar, an ethicist with a background in religious studies, representatives of the public and private insurance sectors, and representatives of the general public" who were tasked with developing recommendations for organ transplantation, including allocation policies); *id.* at 220 (discussing the requirement that procurement organizations include transplant professionals on their board of directors or advisory board); *see also* Organ Procurement and Transplantation Network, 42 C.F.R. § 121.12 (2019) (establishing the Advisory Committee

OPTN's Board of Directors, which develops policies to govern the operation of the OPTN, relies on numerous expert advisory committees.[217] Those committees engage in extensive fact-finding and deliberation.[218]

Since its inception, organ transplantation and allocation have been viewed as both highly technical and deeply political.[219] The algorithmic systems used to allocate organs are provided by the United Network for Organ Sharing (UNOS) under contract with the OPTN, which makes extensive information about the factors and weighting within allocation algorithms available.[220] OPTN/ONUS provides a central source of expertise around algorithmic allocation systems that governs the operation of all member transplant hospitals, organ procurement organizations, and histocompatibility labs in the United States. Thus, while numerous institutions are involved in procuring and transplanting organs, they benefit from a centralized set of expert resources. While bounded by legislation and overseen by the Department of Health & Human Services (HHS),[221] the OPTN/UNOS system for establishing policies to embed in the allocation

---

on Organ Transplantation, governed by the Federal Advisory Committee Act, to provide input on proposed OPTN policies and other matters); 42 C.F.R. § 121.3 (directing The OPTN Board of Directors to include "approximately 50% transplant surgeons or transplant physicians"); Mark D. Stegall et al., *Why Do We Have the Kidney Allocation System We Have Today? A History of the 2014 Kidney Allocation System*, 78 HUM. IMMUNOLOGY 4 (2017) (describing the deliberations and adoption of the kidney allocation system (KAS), including expert input, adopted in December 2014).

217. For a list of current committees, see *Committees*, U.S. DEP'T HEALTH & HUM. SERVS., https://optn.transplant.hrsa.gov/members/committees [https://perma.cc/LWR6-6WWM] (last visited Aug. 1, 2019). Proposed policies OPTN wants to enforce, including allocation policies, must be provided to the Secretary of the Department of Health and Human Services sixty days prior to implementation, the Secretary must publish significant proposed policies in the Federal Register, and they are not enforceable until approved by the Secretary. Department of Health and Human Services. 42 C.F.R. § 121.4 (b)(2) (2019).

218. For a description of a recent process for revising kidney allocations, see Mark D. Stegall et al., *Why Do We Have the Kidney Allocation System We Have Today? A History of the 2014 Kidney Allocation System*, HUM. IMMUNOLOGY 78.1 4, 4–8 (2017). For details on the allocation formula and discussion of use of a software system—Kidney-Pancreas Simulated Allocation Model (KPSAM)—to assess policy proposals, see Bhavna Chopra & Kalathail K. Sureshkumar, *Changing Organ Allocation Policy for Kidney Transplantation in the United States*, 5 WORLD J. TRANSPLANT 38 (2015).

219. *See* Jed Adam Gross, *supra* note 215, at 182–83 (describing issues from blood pressure to distributive justice in congressional hearings).

220. *See Organ Procurement and Transplantation Network Policies*, U.S. DEP'T HEALTH & HUM. SERVS., https://optn.transplant.hrsa.gov/governance/policies/ [https://perma.cc/JC5R-XL4R] (last visited Oct. 11, 2019) (detailing allocation rules and weighting for various organs).

221. The Final Rule requires organ allocation to be based on sound medical judgment; make the best use of donated organs; avoid wasting organs and futile transplants; and promote patient access to transplants. 42 C.F.R. § 121.8 (2019).

algorithms is largely driven by experts enlisted through committees. Those committees are highly specialized, focusing on the specific factors that influence transplant success and considerations of just allocation with respect to particular organs.

The Taxpayer Advocate's office within the IRS offers an alternative model focused on agency development of internal expertise. While not set up specifically to address algorithmic systems, the Taxpayer Advocate has played a key role in identifying the problem with embedded policies in algorithmic systems.[222] Although the Advocate has not been fully successful in ensuring that embedded policies are vetted internally, consistent with other agency review and approval processes for all written policy, procedures, and guidance before issuance and publication, they have drawn agency attention to problematic aspects of systems, some of which have been reformed.[223]

Finally, a bill pending in California takes a different outward-looking approach. Rather than bringing expertise directly into government work, it would leverage the expertise of firms that provide AI-based products and services to public agencies. The bill would require firms to provide information about data curation and processing, as well as bias mitigation strategies, in their contracts with public agencies.[224] Requirements such as this could harness the knowledge and expertise of those close to the problem toward public ends—allowing the experts to identify and address problems, but opening them up for regulator and public scrutiny.[225]

Over the past year, engineers and other employees at firms that build machine learning systems have been increasingly active in objecting to their

222. TAXPAYER ADVOCATE SERV., *supra* note 60 (arguing that "automated systems and software applications require transparency, and employee guidance embedded in systems must be reviewed and continually analyzed for proper application" but noting that "policy guidance embedded in systems is neither reviewed internally nor published externally").

223. *Id.*

224. Artificial Intelligence: Reporting, S.B. 444, 2019 Legis., Reg. Sess. (Cal. 2019), was a state senate bill in California introduced by Senator Umberg (D) on February 21, 2019 to establish that a contractor, vendor, or qualifying business shall maintain a written record of the data used relating to any use of artificial intelligence for the delivery of a product or service to a public entity. The records shall include all the following information: (1) the purpose of the data; (2) a description of the categories of the data; (3) the source of the data; (4) the demographics or information related to a characteristic listed in subdivision (a) of Section 11135 of the Government Code that is used as a source of input data for the creation of the artificial intelligence system. The business shall disclose to public entities that it relies on AI, information about the data, and "its internal policies for how bias in the artificial intelligence system is identified and mitigated." *Id.* at § 3505(c)(3).

225. *See generally* Kaminsky, *supra* note 22, at 26–39 (calling for collaborative governance of algorithms used in the private sector).

use in specific contexts or toward specific ends.[226] Providing engineers and other technical professionals with external justifications for considering the implications of their work on privacy, discrimination, and other substantive values can legitimize these nascent expressions of concern among technical professionals, encourage internal policing,[227] and provide a platform to support collaborative work across institutions on issues such as bias mitigation strategies.

Each of these approaches has their advantages. Given the current level of uncertainty about how best to mitigate bias, leveraging the deep knowledge of the people and institutions close to the problem—as the OPTN and California proposal do in different ways—while exposing them to external review taps into the relevant expertise while fostering attention and accountability to a broader set of values.[228] The Pennsylvania Sentencing Commission, OPTN, and the California proposals lay different internal processes out for the public. OPTN and the Pennsylvania Sentencing Commission provide more detailed information about the properties of the algorithmic systems themselves. Requiring decision makers to explain themselves to outsiders with different perspectives on what policies ought to be embedded, or what bias mitigation strategies should be taken, can push those developing algorithmic systems to engage in more critical technical practice.[229] This approach may yield processes that make agencies, managers, or even individual engineers accountable for considering the impact of algorithmic systems from perspectives beyond speed and task performance, fostering a greater sense of responsibility for the outcomes they produce in the world.

In contrast to the Sentencing Commission and OPTN, the New York City Task Force's mandate cuts across numerous agencies. While the factors contributing to its slow progress are not fully clear, one factor may be that focusing on algorithmic systems writ large, rather than algorithmic systems

226. Meredith Whittaker et al., *AI Now Report 2018*, AI NOW INST. 40–42 (Dec. 2018), https://ainowinstitute.org/AI_Now_2018_Report.pdf [https://perma.cc/7WER-78L4] (summarizing host of employee actions at various companies to stop particular AI projects and system uses).

227. Such approaches could bolster algorithmic whistleblowing by employees envisioned by Sonia Katyal. Katyal, *supra* note 20.

228. *See* KENNETH A. BAMBERGER & DEIRDRE K. MULLIGAN, PRIVACY ON THE GROUND: DRIVING CORPORATE BEHAVIOR IN THE UNITED STATES AND EUROPE 190 (2015) (describing importance of exposing business practices to scrutiny) [hereinafter PRIVACY ON THE GROUND]; *see generally* Kenneth A. Bamberger, *Regulation as Delegation: Private Firms, Decisionmaking, and Accountability in the Administrative State*, 56 DUKE L.J. 377, 445–46 (2006).

229. Philip E. Tetlock, *Accountability: The Neglected Social Context of Judgment and Choice*, 7 RES. ORGANIZATIONAL BEHAV. 297, 314–21 (1985) (reviewing research evidence).

within a particular domain and with participation by substantive experts in that domain, may be perceived as more threatening or less beneficial.

2. A Paradigm for Expert Decision Making

However, these piecemeal attempts to inform case-by-case deliberation with internally- or externally-derived technical expertise about methodologies under consideration offer only a proverbial finger in the dike[230] against the oncoming flood of government usage of machine learning. While helpful, they are fragmentary, costly, and time-consuming. Furthermore, they often (with the possible exception of the New York City Task Force) fail to address the expertise gap across agencies by providing easily deployable, coordinated, and regularized policies and methods governing, and processes for assessing, algorithmic governance tools.

Indeed, centralization of expertise may be particularly important for attending to embedded policies in algorithmic systems. Organ donation is an area where the need for allocation strategies and their deeply political and technical nature were recognized and accounted for up front. But in many areas, algorithmic systems are replacing or aiding decision making that was conceived of and practiced in a more clinical way—relying on human judgment informed by expertise gained through education and training, refined through tacit knowledge and intuition developed experientially through practice, and perhaps discussion and feedback with others—and therefore lacks the level of formalization found in organ donation or even recidivism risk. To the extent algorithmic systems are being introduced in areas where clinical decision-making methods reign, the need for relevant technical expertise in addition to domain-specific expertise will be high; however, internal agency capacity to provide it may be low.

a) The Institutional Paradigm: USDS and the 18F "Skunk Works"

In this vein, efforts begun during the Obama Administration to create centers-of-expertise or shared technical knowledge offer a paradigm for a more systemic approach to the use of technology in government activity. This work has continued with the Office of Federal Procurement Policy (OFPP),[231] which, in conjunction with the USDS, created a Digital IT Acquisition Program (DITAP) to help contracting professionals gain the

230. *See* MARY MAPES DODGE, HANS BRINKER OR THE SILVER SKATES: A STORY OF LIFE IN HOLLAND 105–09 (recounting a tale about a Dutch boy who saves his country by putting his finger in a leaking dike).

231. OFPP was created by Congress to provide overall direction on "Government-wide procurement policies, regulations, procedures, and forms"; and to "promote economy, efficiency, and effectiveness" in procurement. 41 U.S.C. § 1101(b) (2012).

"expertise needed to support the delivery of digital information (i.e., data or content) and transactional services (e.g., online forms and benefits applications) across a variety of platforms, devices, and delivery mechanisms (e.g., websites, mobile applications, and social media)."[232]

President Obama created a centralized pool of technical expertise, USDS, housed within the Executive Office of the President of the United States. USDS is a transitory pool of experts in design, engineering, or product management brought in from outside government for "tours" of service.[233] USDS provides consultation services to federal agencies on information technology and guidance. For example, to develop greater technical expertise among the contracting professionals distributed throughout federal agencies, USDS created the Digital Services Playbook to propagate best practices from both the private and public sector across federal agencies. The Playbook is accompanied by the TechFAR Handbook, which provides guidance to agencies on how to comply with Federal Acquisition Regulations while using the Digital Services Playbook. Similarly, 18F—sometimes described as a skunk works[234] project for government—is an office within the General Services Administration (GSA) that collaborates with other agencies to fix technical problems, build products, and improve how government serves the public through technology.[235] They offer agencies access to expert teams of designers, software engineers, strategists, and product managers skilled in user-centered development, and other design and acquisition expertise.

But in addition to providing experts to government agencies on a task-specific model, 18F also develops guides on topics such as accessibility, agile

232. Memorandum from Lesley A. Field, Deputy Adm'r, Office of Mgmt. & Budget, on Establishment of Federal Acquisition Certification in Contracting Core-Plus Specialization in Digital Services (FAC-C-DS) to Chief Acquisition Officers & Senior Procurement Execs. A-1 (May 18, 2018), https://techfarhub.cio.gov/assets/files/FAC_C_Digital_Services_5-18-18.pdf [https://perma.cc/75MU-WWS9].

233. *How We Work*, U.S. DIGITAL SERV., https://www.usds.gov/how-we-work [https://perma.cc/QZY3-YV6C] (last visited Oct. 3, 2019).

234. *See* Dave Zvenyach, *Joining 18F*, V. DAVID ZVENYACH'S BLOGS (Jan. 20, 2015), https://esq.io/blog/posts/joining-18f/ [https://perma.cc/VM9T-F4C8] (describing 18F as "a modern-day digital skunk works . . . [housing] some of the best and brightest developers in America, building applications that agencies need in a modern, experimental, and explicitly iterative manner"); *accord* Jason Bloomberg, *Digital Influencer Jez Humble: DevOps For 'Big Hairy Enterprises'*, FORBES (Mar. 31, 2016, 9:54 AM), https://www.forbes.com/sites/jasonbloomberg/2016/03/31/digital-influencer-jez-humble-devops-for-big-hairy-enterprises/#2f603f054a21 [https://perma.cc/8B3G-C8R4] (describing 18F as "a skunkworks-like team of designers, developers, and product specialists").

235. *About*, 18F, https://18f.gsa.gov/about/#our-team [https://perma.cc/2DT4-K9FQ] (last visited Oct. 3, 2019).

development, and design methods to assist federal agencies.[236] An example of such guidance documents is the U.S. Web Design Standards (USWDS),[237] a joint product of 18F and USDS.[238] The USWDS provide guidance, templates, and models for developers and designers; it covers design including accessibility, front end and back end coding, and provides code and performance guidelines.[239] The mix of centralized expert staff who can be deployed for limited periods of time to assist agencies with complex technology projects and detailed guidance documents is appealing in the context of algorithmic systems.

Coordinated expert input through shared resources in the vein of the USWDS should frame agency decision making around algorithmic systems, as well as consulting services to provide hands-on assistance to agencies on an as-needed basis. The New York City Task Force was an effort to develop a methodology for eliciting relevant information from developers and operators of systems through targeted questions. It builds on research in the FAT* community identifying specific information necessary to understand appropriate uses of and potential biases in systems.[240] The recently introduced Algorithmic Accountability Act of 2019,[241] while aimed at the private sector rather than the public sector, takes a similar approach by requiring automated decision system, data protection impact assessments, and regulations promulgated by the Federal Trade Commission to guide such assessments. A similar mix of standardized questions, guidance, and localized

236. *Guides*, 18F, https://18f.gsa.gov/guides/ [https://perma.cc/N2SD-QY7B] (last visited Oct. 3, 2019).

237. U.S. WEB DESIGN SYS., https://designsystem.digital.gov/ [https://perma.cc/2V2Y-K2HN] (last visited Oct. 3, 2019).

238. Mollie Ruskin et al., *Introducing the U.S. Web Design Standards*, 18F (Sept. 28, 2015), https://18f.gsa.gov/2015/09/28/web-design-standards/ [https://perma.cc/X33R-P43C].

239. U.S. WEB DESIGN SYS., *supra* note 237.

240. *See* Margaret Mitchell et al., *Model Cards for Model Reporting*, PROC. CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 220 (2019) (proposing information about what the use context model was designed for, model performance benchmarked across different groups, and processes of validation and accompany models); *see also* Galen Harrison et al., *Towards Supporting and Documenting Algorithmic Fairness in the Data Science Workflow*, WORKSHOP ON TECH. & CONSUMER PROTECTION (May 23, 2019), https://www.ieee-security.org/TC/SPW2019/ConPro/papers/harrison-conpro19.pdf [https://perma.cc/VPZ9-B4UN] (proposing documentation and visualization of algorithms in data science processes); *see generally* Timnit Gebru et al., *Datasheets for Datasets*, ARXIV (Apr. 16, 2019), https://arxiv.org/pdf/1803.09010.pdf [https://perma.cc/82DT-G93P] (proposing information about attributes of datasets that should be documented and shared). These efforts resemble the efforts in the reproducible research community to provide information about data, code, computational steps, software environment, etc. *See* Victoria Stodden et al., *Enhancing Reproducibility for Computational Methods*, 354 SCIENCE 1240 (2016).

241. S. 1108, 116th Cong. (2019); H.R. 2231, 116th Cong. (2019) (providing the companion House bill).

assessment is found in the privacy area, where Congress required administrative agencies to conduct privacy impact assessments[242] but authorized the Office of Management and Budget (OMB) to issue detailed guidance for agency implementation.[243]

The mix of expertise offered by 18F and USDS, and their track record of providing effective guidance and leadership, provides the most compelling model to begin developing methods and tools for cross-agency efforts to identify and reason about embedded policies in algorithmic decision-making systems.

b) Models to Inform the Centralized Process

Existing legal frameworks addressing the use of predictive models in the area of credit and employment offer guidance regarding the possible content of centralized processes. For example, under the Equal Credit Opportunity Act (ECOA), in order to use age[244] as a predictive factor in granting credit, creditors must use an "empirically derived, demonstrably and statistically sound, credit scoring system."[245] To meet these criteria, the system must be:

> (i) [b]ased on data that are derived from an empirical comparison of sample groups or the population of creditworthy and non-creditworthy applicants who applied for credit within a reasonable preceding period of time;
>
> (ii) [d]eveloped for the purpose of evaluating the creditworthiness of applicants with respect to the legitimate business interests of the

242. E-Government Act of 2002, Pub. L. No. 107-347, 116 Stat. 2899, 2921 (2002) (requiring agencies to conduct a privacy impact assessment before "developing or procuring information technology that collects, maintains, or disseminates information that is in an identifiable form").

243. *See* Memorandum from Joshua B. Bolten, Dir., Office of Mgmt. & Budget, on OMB Guidance for Implementing the Privacy Provisions of the E-Government Act of 2002 to Heads of Exec. Dep'ts & Agencies (Sept. 26, 2003), https://obama whitehouse.archives.gov/omb/memoranda_m03-22/ [https://perma.cc/39ZE-U2TC].

244. While these requirements only apply to the use of credit scoring systems that use age as a predictive factor in granting credit, regulator comments suggest that in practice, statistical validation of credit scoring systems is a key tool that the Federal Trade Commission and Consumer Financial Protection Board use to assess compliance with ECOA and that regulated entities meet these guidelines to manage risk. *See Credit Scoring: Testimony Before the U.S. House Subcomm. on Financial Institutions & Consumer Credit*, 111th Cong. (Mar. 24, 2010), https://www.federalreserve.gov/newsevents/testimony/braunstein 20100323a.htm [https://perma.cc/ZU98-N33W] (testimony of Sandra F. Braunstein, Dir., Div. of Consumer & Cmty. Affairs, Fed. Reserve).

245. *See* 12 C.F.R. § 1002.2(p) (defining empirically-derived and other credit scoring systems); 12 C.F.R. app. I § 1002.2(p) (2019) ("1. . . . The definition under §§ 1002.2(p)(1)(i) through (iv) sets the criteria that a credit system must meet in order to use age as a predictive factor.").

> creditor utilizing the system (including, but not limited to, minimizing bad debt losses and operating expenses in accordance with the creditor's business judgment);
>
> (iii) [d]eveloped and validated using accepted statistical principles and methodology; and
>
> (iv) [p]eriodically revalidated by the use of appropriate statistical principles and methodology and adjusted as necessary to maintain predictive ability.[246]

While designed to regulate credit granting, these criteria, along with the research on aspects of models and data sets necessary to determine appropriate and fair uses, provide a useful starting point for thinking about data and models.

The Uniform Guidelines on Employee Selection Procedures (Uniform Guidelines)[247] issued in 1978 by the Equal Employment Opportunity Commission, provides another set of criteria that might be useful in the context of developing guidance on algorithmic systems. These guidelines are designed "to assist employers, labor organizations, employment agencies, and licensing and certification boards to comply with requirements of Federal law prohibiting employment practices which discriminate on grounds of race, color, religion, sex, and national origin."[248] The Uniform Guidelines specify that selection procedures having an adverse impact on these protected characteristics will be prohibited:

> The use of any selection procedure which has an adverse impact on the hiring, promotion, or other employment or membership opportunities of members of any race, sex, or ethnic group will be considered to be discriminatory and inconsistent with these guidelines, unless the procedure has been validated in accordance with these guidelines, or the provisions of section 6 below are satisfied.[249]

The guidelines thus provide a framework for determining the proper use of tests and other selection procedures. Moreover, while they do not require employers to conduct validity studies where no adverse impact results, they draw attention to the potential disparate impacts of selection procedures by providing guidance on how to construct and validate them—encouraging more thoughtful technical choices. For example, the Uniform Guidelines state that "where two or more selection procedures are available which serve

246. 12 C.F.R. § 1002.2(p), n.217.
247. 29 C.F.R. § 1607 (2019).
248. 29 C.F.R. § 1607.1.
249. 29 C.F.R. § 1607.3.

the user's legitimate interest . . . and which are substantially equally valid for a given purpose, the user should use the procedure which has been demonstrated to have the lesser adverse impact."[250] While these guidelines already apply to the employment practices of agencies, they are equally applicable to the development of algorithmic systems more generally and could easily be adapted for such purposes. Other elements of the Uniform Guidelines are similarly relevant, such as the discussion of criterion-related, content, and construct validity studies to assess selection procedures.[251] In particular, the Uniform Guidelines clarify that the validity of a selection procedure can be validated by empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance; by data showing that the content of the selection procedure is representative of important aspects of performance on the job; or through a construct validity study that consists of data showing that the procedure measures the degree to which candidates have identifiable characteristics which have been determined to be important in successful performance in the job.[252] In particular, the distinction between criterion-related validity and construct validity could be used to clarify different ways in which algorithmic systems could be reviewed. Although the Uniform Guidelines were issued thirty years ago, their focus on the *impact* of selection procedures—procedures used to predict how an applicant will do at a job—continues to make them relevant in the context of automated decisions today.

In conjunction with the growing set of guidance documents, toolkits, and other methods for analyzing algorithmic systems, these existing government regulations could provide a starting point for developing a suite of guidance documents and methods to standardize the questions and processes federal agencies use to assess and design algorithmic systems. With the availability of appropriate experts at 18F and USDS, this could provide a flexible, on-demand set of tools and personnel to fill the expertise gap plaguing federal agencies.

B. INFUSING AGENCY DELIBERATION WITH POLITICAL VISIBILITY

1. Impact Assessments: Bridging Technocracy and Democracy in Agency Deliberation

Algorithmic impact assessments provide a critical tool to bridge the technocratic and democratic elements of deliberation.[253] Such tools not only

250. 29 C.F.R. § 1607.3.B.
251. *See, e.g.*, 29 C.F.R. § 1607.5.B.
252. *Id.*
253. *See* Reisman et al., *supra* note 22; Selbst, *supra* note 22.

enable and trigger agency deliberation about the technical aspects of system design, but surface the political implications of those choices, offering an important prerequisite for reasoned consideration of the policies they embed by agency staff, the public, and the political branches. They thus bridge the dual deliberation requirements of substantive expertise and political visibility.

More specifically, algorithmic impact assessments are boundary negotiation objects—objects used to "record, organize, explore and share ideas; introduce concepts and techniques; create alliances; create a venue for the exchange of information; augment brokering activities; and create shared understanding."[254] If publicly disclosed, algorithmic impact assessments mediate between experts and the public, providing a common reference point, but importantly do not reflect a common understanding or consensus across these groups.[255] A good impact assessment provides a tool for exploring the problem space, helping the public collectively consider the points of policy within a machine learning system. They also allow the public and expert communities to in effect argue about the boundary between science and policy; in this way, they facilitate negotiation not just across the boundary, but about the location of it.

It is no accident that many of the "arbitrary and capricious" cases finding that agency action fails to reflect appropriate deliberation about relevant analytic techniques (models, assumptions, the use of data, and choices between false-negatives and false-positives) arise in the review of either cost-benefit or environmental-impact statements—two types of impact assessments well-enshrined in administrative law. Such tools are instrumental in "making bureaucracies think,"[256] and "take a hard look at the potential . . . consequences of their actions."[257] While they do not mandate substantive

254. Matthew J. Bietz & Charlotte P. Lee, *Collaboration in Metagenomics: Sequence Databases and the Organization of Scientific Work*, PROC. 11TH EUR. CONF. ON COMPUTER SUPPORTED COOPERATIVE WORK 243, 247.

255. (2009) Privacy regulators and scholars generally advocate for publishing privacy impact assessments to support "contestation and public debate." Jennifer Stoddart, *Auditing Privacy Impact Assessments: The Canadian Experience*, *in* PRIVACY IMPACT ASSESSMENT 453–54 (David Wright & Paul De Hert eds., 2012) (describing various regulators' positions on publication). Unfortunately, proposed bills in the United States do not require publication. *See* The Algorithmic Accountability Act of 2019, S. 1108, 116th Cong. (2019); H.R. 2231, 116th Cong. (2019); *see also* Margot E. Kaminski & Andrew D. Selbst, *The Legislation That Targets the Racist Impacts of Tech*, N.Y. TIMES (May 7, 2019), https://www.nytimes.com/2019/05/07/opinion/tech-racism-algorithms.html [https://perma.cc/P6YV-43AV] (critiquing the lack of a publication requirement).

256. SERGE TAYLOR, MAKING BUREAUCRACIES THINK: THE ENVIRONMENTAL IMPACT STATEMENT STRATEGY OF ADMINISTRATIVE REFORM 251 (1984).

257. *The National Environmental Policy Act: A Study of Its Effectiveness After Twenty-Five Years*, COUNCIL ON ENVTL. QUALITY iii (Jan. 1997), https://ceq.doe.gov/docs/ceq-publications/nepa25fn.pdf [https://perma.cc/4KJH-4VK5] (discussing the National

outcomes, they force administrative agencies to turn their analytic capacity towards particular issues, and require explicit and publicly reviewable identification, recognition, and explanation of their choices about them.

Such requirements are especially important in surfacing issues in contexts with which government actors are unfamiliar, or where the issues addressed are "orthogonal to" or even "in tension with, an agency's primary mission."[258] Our research has demonstrated the particular importance of such impact assessments when agencies engage with the use of technology, given problems of technical illiteracy, design opacity, and the phenomena by which the modality "hide[s] the political nature of its design,"[259] rendering policy implications invisible and making choices seem fixed, natural, and incontestable. In particular, the requirement imposed by the E-Government Act of 2002,[260] that agencies complete privacy impact assessments (PIAs) when developing or procuring information technology systems that include personally identifiable information, has forced agencies to address important privacy implications of systems intended to promote a range of public aims—implications that remained unnoticed and unaddressed (with important and expensive security consequences) where such requirements were not satisfied robustly.[261]

When agencies adopt technology, we have argued accordingly, the choices that impact legal rights must be addressed through the use of impact assessment tools.[262] On the one hand, those tools "create different frameworks and bring new considerations to bear in agency actions," as well as bridge the gulf between the substantive domain expertise of agency staff and the frameworks and knowledge of outside experts.[263] On the other, they "facilitate participation by issue experts and by stakeholders who might otherwise be unaware of relevant risks and technological alternatives."[264]

Environmental Policy Act's "success" in making federal agencies take a "hard look" at the potential environmental consequences of their actions).

258. Kenneth A. Bamberger & Deirdre K. Mulligan, *Privacy Decisionmaking in Administrative Agencies*, 75 U. CHI. L. REV. 75, 83 (2008).

259. Mulligan & Bamberger, *supra* note 12, at 778.

260. E-Government Act of 2002, Pub. L. No. 107-347, 116 Stat. 2899, 2921 (2002).

261. *See* Kenneth A. Bamberger & Deirdre K. Mulligan, *PIA Requirements and Privacy Decision-Making in US Government Agencies*, *in* PRIVACY IMPACT ASSESSMENT 225–50 (David Wright & Paul De Hert eds., 2012) (comparing DHS's responsible adoption of RFID technology with that of the Department of State, which failed to discuss technical aspects of the program, alternative technologies, and risks); Bamberger & Mulligan, *supra* note 258 (discussing the same).

262. Mulligan & Bamberger, *supra* note 12, at 764–66, 780 (arguing that when agencies govern "by design," they should use values-impact assessment tools, such as a "human rights impact assessment").

263. *Id.* at 765.

264. *Id.*

To have the desired ameliorative effect, experts must be involved in conducting impact assessments[265] and the outputs must be available to the public.[266] Current and proposed efforts to require government agencies to conduct impacts of surveillance and algorithmic systems too often emphasize process while shorting or overlooking expertise.[267] There is an ongoing rise in chief privacy officers and other privacy staff, and more recently a move to redefine their roles to provide them greater latitude to address harms that may flow from data analysis and applications—reflected in new titles such as chief data governance officer or information steward.[268] More recently, growing concerns with AI—ranging from job displacement to bias to military applications—have ushered in a new set of professional staff with titles such as chief ethics officer.[269] More significantly, a major shift in privacy regulation in the European Union was to require data protection officers. The emphasis on professionals reflects the growing understanding that particular expertise is necessary to fully use tools such as impact assessments.

2. Other Political Visibility-Enhancing Processes

The impact assessments described above provide foundational tools for spanning boundaries between expert communities, policy makers, and the

265. Bamberger & Mulligan, *supra* note 258, at 104 (concluding that optimal use of privacy impact assessment turned on expert inter-disciplinary staff).

266. While different countries have taken different positions on whether privacy impact assessments or summaries ought to be shared with the public, there is a general preference for doing so due to the recognition that doing so can maintain public trust and confidence in systems and organizations. *See* David Wright & Paul De Hert, *Introduction to Privacy Impact Assessment*, *in* PRIVACY IMPACT ASSESSMENT 3–32 (David Wright & Paul De Hert eds., 2012); *see also* Elin Palm & Sven Ove Hansson, *The Case for Ethical Technology Assessment (eTA)*, 73 TECHNOLOGICAL FORECASTING & SOC. CHANGE 543, 547 (2006) (arguing for publication because "[i]t would be delusive to believe that technology developers are conscious of all the effects of their products[; i]n many cases, negative side effects come as a surprise to technology developers themselves").

267. *See* OAKLAND, CAL., MUN. CODE §§ 9.64.010–9.64.070 (2018) (establishing requirement for impact assessments but not requiring new staff); *accord* SANTA CLARA CITY, CAL., ORDINANCE CODE § A40 (2016).

268. *See, e.g.*, Molly Hulefeld, *What is a chief data ethics officer anyway*, IAPP PRIVACY ADVISORY (Nov. 27, 2018), https://iapp.org/news/a/making-way-for-the-rise-of-the-chief-data-ethics-officer/ [https://perma.cc/Q39K-WDUV] ("Acxiom renamed its privacy program to become the[']data ethics, governance, protection and privacy program.' "); *Mastercard Names JoAnn Stonier Chief Data Officer*, MASTERCARD (Feb. 8, 2018), https://newsroom.mastercard.com/press-releases/mastercard-names-joann-stonier-chief-data-officer/ [https://perma.cc/BT43-QVX6] (announcing that JoAnn Stonier was moving from chief information governance and chief privacy officer to chief data officer, a new position designed to affirm the company's commitment to data protection).

269. *See, e.g.*, *Rise of The Chief Ethics Officer*, FORBES (Mar. 27, 2019), https://www.forbes.com/sites/insights-intelai/2019/03/27/rise-of-the-chief-ethics-officer/#5e67f2ba5aba [https://perma.cc/8P84-HFYR].

general public. By distilling the policy-relevant choices in technical design, impact assessments surface issues for political consideration.

However, the question of what aspects of a technical system are political is itself a value judgment, and who decides is itself a political matter. Despite the best efforts of experts,[270] the political and the technical defy clean separation.[271] The complexity and density of technical and scientific matters can create barriers to broader participation in policy debates. This is surely true with respect to machine learning systems, where the complex interaction of design choices, data, and interfaces can produce clearly political outcomes yet be shielded from public scrutiny. Thus, agencies must further employ a broader set of processes to publicize system politics and elicit the public participation essential for legitimate administrative decision making.

The COMPAS debates provide a glimpse of the politics of expertise in action in machine learning systems. After journalists at ProPublica exposed the different false positive and false negative rates for black and white defendants, Northpointe, the developer of COMPAS, defended the system because it was "equally accurate for blacks and whites," asserted its status as expert on the matter, and dismissed ProPublica's analysis as wrong because they failed to "account the different base rates of recidivism for blacks and whites."[272] Their rejoinder attempts to remove a legitimate political question—should models of fairness account for different base rates—from the public discussion by framing the question as one of whether to account for objective facts. Academics later pointed out that the differing perspectives on how to conceive of fairness espoused by Northpointe and ProPublica were mathematically incompatible, yet both defensible and possibly mutually desirable.[273] Here, as in many other instances, the technical

270. Jasanoff has documented how experts in science policy constantly attempted to demarcate the politics from the scientific in their practice in an effort to preserve claims of objectivity. *See* Sheila S. Jasanoff, *Contested Boundaries in Policy-Relevant Science*, 17 SOC. STUDS. SCI. 195, 199 (1987) ("To shore up their claims to cognitive authority, scientists have to impose their own boundaries between science and policy.").

271. *See generally* STATES OF KNOWLEDGE: THE CO-PRODUCTION OF SCIENCE AND SOCIAL ORDER (Sheila Jasanoff ed., 2004) (presenting a collection of essays exploring the ways in which our methods of understanding and reasoning about the world and choices about how to live in the world are interdependent).

272. William Dieterich et al., *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, NORTHPOINTE 9 (July 8, 2016), http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf [https://perma.cc/WQQ5-SEMC].

273. *See* Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear*, WASH. POST (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.ce67b2c3fa95 [https://perma.cc/5TDR-E4MQ]; *see also* Jon Kleinberg et al., *Inherent Trade-*

and the political are entangled. Yet the important question about what metric should be used to support fair risk assessments appears to have escaped careful scrutiny by the public agency.

Fostering public engagement requires processes at a variety of levels, and times, to surface the politically salient questions latent in the design and use of machine learning systems. As an initial matter, public engagement would *both* be primed at the abstract level—i.e., the public would have a general understanding of the important policy choices entangled with system design—*and* invited to participate in discourse with respect to specific systems.

a) Fostering Ongoing Public Engagement Through Agenda-Setting

Today, the politics of machine learning systems are an object of ongoing public scrutiny. The "black boxes" are being aired out, and centers of science policy have participated in opening up algorithmic systems to scrutiny. During the Obama Administration, the Office of Science and Technology Policy wrote several reports exploring the politics of the design and use of big data and artificial intelligence.[274] In particular, the reports detailed the biases that can result from training data and model design.[275] The sustained attention to the political issues embedded in technical systems was particularly important given President Obama's efforts to strengthen "America's role as the world's engine of scientific discovery and technological innovation" and use technology to improve service delivery

Offs In The Fair Determination of Risk Scores, ARXIV (Nov. 17, 2016), https://arxiv.org/pdf/1609.05807.pdf [https://perma.cc/MR4S-9W8W].

274. *See, e.g.*, *Big Data: Seizing Opportunities, Preserving Values*, EXECUTIVE OFF. PRESIDENT (May 2014), https://hsdl.org/?view&did=752636 [https://perma.cc/AV9Y-L4FR]; *Big Data and Differential Processing*, EXECUTIVE OFF. PRESIDENT (Feb. 2015), https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/docs/Big_Data_Report_Nonembargo_v2.pdf [https://perma.cc/Z9AL-48C7]; *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, EXECUTIVE OFF. PRESIDENT (May 2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf [https://perma.cc/VJ3S-HW3G]; Ed Felten & Terah Lyons, *The Administration's Report on the Future of Artificial Intelligence*, WHITE HOUSE BLOG (Oct. 12, 2016, 6:02 AM), https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence [https://perma.cc/X4EQ-CFWT]; *see also* John P. Holdren & Megan Smith, *Cabinet Exit Memo*, OFF. SCI. & TECH. POL'Y, EXECUTIVE OFF. PRESIDENT (Jan. 5, 2017), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_exit_memo_final.pdf [https://perma.cc/FN5W-7S6B].

275. *See Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, *supra* note 274.

and governance across the public sector.[276] The Federal Trade Commission conducted workshops and reports that similarly focused attention on the policy implications of algorithmic systems, but with attention to its implications in the marketplace rather than the civic sector.[277] These reports focused agencies and the public on the politics of algorithmic choices, and sent a strong political signal to agencies that technological design ought to be scrutinized as policy.

Agenda-setting activities such as these White House reports, other Administration strategy documents that address the ethical, legal, and social aspects of artificial intelligence,[278] and studies and reports by organizations such as the National Academy of Science,[279] throw open the doors to ongoing public engagement through plain language, case studies, context, and explicit identification of the policy judgments that warrant public engagement along with technical expertise. They also encourage agencies to surface questions about specific technical system designs for public scrutiny.

b) Fostering Public Engagement on Specific Systems

Broader participation in the adoption and design of specific systems poses some practical challenges. As discussed in Part II, unlike regulations which agencies write themselves, the software code of machine learning systems is typically authored and owned by companies. Empirical work suggests that agencies routinely have little impact on the design of such systems, largely procuring them "off the shelf." Where an agency desires to engage the public with the design of a technical system, contracts[280] and

276. Holdren & Smith, *supra* note 274, at 2 & n.2 (citing President Obama's remarks on November 23, 2009 at the launch of his "Educate to Innovate" campaign for excellence in science, technology, engineering, and math education).

277. *See Big Data: A Tool for Inclusion Or Exclusion?*, FED. TRADE COMMISSION (Sept. 15, 2014), https://ftc.gov/news-events/events-calendar/2014/09/big-data-tool-inclusion-or-exclusion [https://perma.cc/69PE-4ZGH]; *Data Brokers: A Call for Transparency and Accountability*, FED. TRADE COMMISSION (May 2014), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf [https://perma.cc/6QSU-SD8A].

278. *See The National Artificial Intelligence Research and Development Strategic Plan*, NAT'L SCI. & TECH. COUNCIL, at 8–40 (Oct. 2016), https://nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf [https://perma.cc/JBV5-Z5HK].

279. *See, e.g.*, NAT'L ACADS. SCIS., ENG'G & MED., PROACTIVE POLICING: EFFECTS ON CRIME AND COMMUNITIES (David Weisburd & Malay K Majmundar eds., 2018), https://doi.org/10.17226/24928 [https://perma.cc/7A3H-USQV] (discussing biases in statistical predictions).

280. *See* Joseph Lorenzo Hall, *Contractual Barriers to Transparency in Electronic Voting*, PROC. 2007 USENIX/ACCURATE ELECTRONIC VOTING TECH. WORKSHOP (2007), https://www.usenix.org/legacy/event/evt07/tech/full_papers/hall/hall_html/jhall_evt07_

intellectual property[281] may present formidable obstacles. Testing—a tried and true method of exploring how a system performs for various populations or under different conditions (usability) as well as security properties—may be contractually limited.[282] Companies may even demand secrecy of training materials. While revealing source code may not be a necessary or desirable means of exposing the values embedded in systems, allowing regulators and experts of their choosing to examine, test, and tinker with systems is an initial prerequisite for surfacing values for public consideration.

As noted elsewhere, public participation is important during both design and deployment.[283] The U.S. Public Participation Playbook provides best practices for agencies to engage the public. While the strategies it has identified through collaboration with agency staff and citizen engagement experts are technology agnostic, they provide guidelines for and examples of successful citizen engagement. Of particular importance for machine learning is the Playbook's mindset of ongoing public engagement and feedback, its emphasis on identifying engagement strategies for specific stakeholder groups, and its emphasis on using prototypes and systems as well as more traditional means of eliciting feedback.

Combining the high-level airing of issues in the White House reports discussed above with participatory processes at the USDS and 18F could provide a powerful first step in a framework for public engagement with the politics of machine learning systems. The development of the Web Standards by USDS and 18F was done through an open process and produced publicly available resources for both use and interrogation. In the words of USDS and 18F, "we're working in the open to create a resource that everyone can own and contribute to."[284] Admittedly, this sort of openness is not conducive to broad participation by the lay public, but it does open up technical design for interrogation by experts who can support such participation and deliberation.

However, USDS and 18F go beyond such technocratic openness. For example, 18F played a key role in implementing the Digital Accountability

---

html.html [https://perma.cc/GUY2-HY4X] (discussing the use of contracts to limit public oversight).

281. *See, e.g.*, Aaron Burstein et al., LEGAL ISSUES FACING ELECTION OFFICIALS IN AN ELECTRONIC-VOTING WORLD (Mar. 15, 2007), https://www.law.berkeley.edu/files/Legal_Issues_Facing_Election_Officials.pdf [https://perma.cc/7DKX-JPG8]; Levine, *supra* note 114.

282. Hall, *supra* note 280.

283. *See* Mulligan & Bamberger, *supra* note 12, at 772 (arguing that "[t]he traditional sequential perspective of 'policymaking' (during which there is an opportunity for input) followed by 'implementation' is inconsistent with [regulation through] design").

284. Ruskin, *supra* note 238.

and Transparency Act,[285] building a prototype of the proposed technical implementation to facilitate public feedback. In effect, through prototyping, 18F brought public participation into the agile software design process.[286] The prototyping used in this instance was one of many strategies for fostering and improving public participation in shaping government programs, including regulations, developed by 18F in conjunction with numerous government agencies.

Yet the broader public must be brought in, as well. The Federal Trade Commission's practices used to develop policy around the privacy impact of technology use and design model suggest two important elements of agency process: publicity and engagement.[287] The Commission conducted and publicized research on both child-directed and general audience websites, increasing the transparency of technology privacy behavior and its policy implications, spurring public attention and providing the basis for engagement. The participatory fora that followed empowered privacy advocates (indeed, many advocacy organizations were founded in response to this opportunity) who both provided input and served as a means for publicizing policy threats more broadly.

Meaningful public participation focused more directly on the adoption of specific systems requires similar scaffolding by agencies. Prototypes and simulations can be powerful means of publicizing technology choices by translating between mathematical formulations and policy choices. For example, the debate about how to model fairness in COMPAS (described above) may seem esoteric to members of the public, but a simulation that allowed individuals to play with various fairness metrics could powerfully expose them to the political implications. An example of such a tool is Google's What-if Tool. The What-if Tool is an interactive visual interface that allows individuals to probe machine learning models. The tool provides options to explore various algorithmic fairness constraints, compare counterfactuals, selectively add and remove data points, compare models on a single data set, among others.[288] Interactive exploratory tools provide opportunities for experiential learning by the public, building greater understanding of aspects of model design, and potentially framing questions about specific model design or data selection choices facing an agency.

285. Digital Accountability and Transparency Act of 2014, Pub. L. No. 113-101, 128 Stat. 1146.

286. *See Implementing A Government-wide Law*, 18F, https://18f.gsa.gov/what-we-deliver/data-act/ [https://perma.cc/WAC6-7FKF] (last visited Oct. 9, 2019).

287. PRIVACY ON THE GROUND, *supra* note 228, at 189–90 (discussing FTC efforts).

288. WHAT-IF TOOL, https://pair-code.github.io/what-if-tool/ [https://perma.cc/HVJ6-HHWN] (last visited Oct. 9, 2019).

The models of the Oakland Privacy Advisory Committee, the New York City Algorithmic Taskforce, and the Pennsylvania Sentencing Commission, moreover, use public events in which experts engage with impact assessments and systems before the public to further elaborate the policies that warrant attention. In effect, these models bolster the technological expertise of the public. Given that civil society organizations often have limited technical capacity, government provisioning of such experts to the public and stakeholders through expert committees may be an important component of the public participation infrastructure. Without agency enlistment of such expertise for the public, the public may miss risks and opportunities posed by machine learning systems and be unable to formulate appropriate and viable solutions.

3. Contestable Design

To ensure that the informed agency engages with and deliberates about the policymaking that occurs through system deployment and use, government machine learning systems must be designed to promote contestability.[289] That is, they must be designed to reveal their "thinking" and receive feedback from and collaborate with human users at runtime.[290] By fostering user engagement within the system, contestable systems use that engagement to iteratively identify and embed domain knowledge and contextual values as decision making becomes a collaborative effort in sociotechnical systems. Contestable systems thus provide a means for system logic to be overseen by, to learn from, and to be shaped by, the domain

289. *See* Tad Hirsch et al., *Designing Contestability: Interaction Design, Machine Learning, and Mental Health*, 2017 PROC. ACM DESIGNING INTERACTIVE SYSS. CONF. 95, 98 (2017) (identifying "contestability" as a new design concern—focused on anticipating and designing for the ways technology can reshape knowledge production and power and describing three lower-level design principles to support contestability: 1) improving accuracy through phased and iterative deployment with expert users in environments that encourage feedback; 2) heightening legibility through mechanisms that "unpack aggregate measures" and "trac[e] system predictions all the way down" so that "users can follow, and if necessary, contest the reasoning behind each prediction"; and relatedly, in an effort to identify and vigilantly prevent system misuse and implicit bias, 3) identifying "aggregate effects" that may imperil vulnerable users through mechanisms that allow "users to ask questions and record disagreements with system behavior" and engage the system in self-monitoring).

290. Contestable design aligns with Mireille Hildebrandt's call for " 'agonistic machine learning,' i.e., demanding that companies or governments that base decisions on machine learning must explore and enable alternative ways of datafying and modelling the same event, person or action." Mireille Hildebrandt, *Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning*, 20 THEORETICAL INQUIRIES L. 83, 106 (2019). This "highlights the contestability at the level of the inner workings of machine learning systems." *Id.* at 118. Also, this "responds to the need to call out the ethical and political implications of *who decides task T, performance metric P and experience E, and to investigate how this is done, taking into account which (and whose) concerns are at stake*." *Id.* at 110.

expertise and experience of human agency staff actually vested with administrative discretion. Contestability is thus one way "to enable responsibility in knowing," to use Judith Simon's phrase, as the production of knowledge is spread across humans and machines.[291]

As legal scholars, we are reluctant to argue for generalized design choices for agency machine learning systems. However, existing literature suggests that particular design choices support contestability by fostering user understanding of models and outputs, collaborative construction of systems, dynamic feedback and control, and within-model challenges to system outputs.[292] These approaches suggest the beginnings of a list of design requirements that permit contestability, which frames an agenda for research and experimentation about governmental systems looking forward.

---

291. Judith Simon, *Distributed Epistemic Responsibility in a Hyperconnected Era*, *in* THE ONLIFE MANIFESTO: BEING HUMAN IN A HYPERCONNECTED ERA 145–59, 146 (Luciano Floridi ed., 2015), https://doi.org/10.1007/978-3-319-04093-6_17 [https://perma.cc/2BQV-DBEG].

292. For insights on how contestable systems advance individual understanding, see, for example, Motahhare Eslami, *Understanding and Designing Around Users' Interaction with Hidden Algorithms in Sociotechnical Systems*, 2017 CSCW '17 COMPANION 57 (describing several studies finding that seamful designs, which expose algorithmic reasoning to users, facilitated understanding, improved user engagement, and in some instances altered user behavior); Motahhare Eslami et al., *"I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds*, 2015 PROC. 33RD ANN. ACM CONF. ON HUM. FACTORS COMPUTING SYSS. 153 (describing the lasting effects on how users engage with Facebook to influence the News Feed algorithm after an experimental design intervention that visualized its curatorial voice); Susan Joslyn & Jared LeClerc, *Decisions with Uncertainty: The Glass Half Full*, 22 CURRENT DIRECTIONS PSYCHOL. SCI. 308 (2013) (describing how displaying uncertainty in weather predictions can lead to more optimal decision making and trust in a forecast: transparency about probabilistic nature of prediction engenders trust even when predictions are wrong); Simone Stumpf et al., *Toward Harnessing User Feedback For Machine Learning*, 2007 PROC. 12TH INT'L CONF. ON INTELLIGENT USER INTERFACES 82; Simone Stumpf et al., *Interacting Meaningfully with Machine-Learning Systems: Three Experiments*, 67 INT'L J. HUMAN-COMPUTER STUD. 639 (2009) (noting that explainable systems can improve user understanding and use of system and enable users to provide deep and useful feedback to improve algorithms); Travis Moor et al., *End-User Debugging of Machine-Learned Programs: Toward Principles for Baring the Logic*, SEMANTIC SCHOLAR (2009), https://pdfs.semanticscholar.org/9a4f/a9f116668927575113d6d2e8572e39925650.pdf?_ga=2.190583149.793734117.1566762743-966572889.1566762743&_gac=1.216591076.1566762743.CjwKCAjw44jrBRAHEiwAZ9igKEyD8gmFQSr6pK9M8nWVxjHzez0odTFCvRLu4dS7rXCjRvZGxRwsmBoCVXwQAvD_BwE [https://perma.cc/E7BA-MXS3] (noting that salient explanations helped users adjust their mental models); Saleema Amershi et al., *Power to The People: The Role of Humans in Interactive Machine Learning*, 35 AI MAG. 105 (2014) (providing an overview of interactive machine learning research with case studies, and discussing value of interactive machine learning approaches for the machine learning community as well as users).

a) Design Should Expose Built-in Values

At its core, contestability requires systems to be designed in a way that exposes value-laden features and parameters. Most simply, this visibility can prompt awareness, reflection, and feedback by agency decision makers relying on these systems. Reminding agency staff about the original choice of model and training data can prompt future deliberation about their appropriateness, particularly where a system consistently produces outcomes that agency staff perceive as incorrect in certain classes of cases. Decision makers relying on recidivism risk systems, such as Northpointe's COMPAS, must be made fully aware of how gender and other protected attributes are used, to both educate them about the policy choices underlying their decision supports and to encourage feedback from ground-level staff about their strengths and limitations. Brauneis and Goodman underscore the ways that contestability can foster ongoing deliberation about policy, describing the ways that the Hunchlab predictive policing software

> allows each community to set weights for the relative seriousness of each type of crime—how much more important is it to stop a murder than a burglary? It also allows tailored weights for patrol efficacy—[for example,] indoor crimes are less likely to be deterred by increased police presence.[293]

Moreover, design decision visibility can enable active participation by system users in consequential decisions about their configuration or use. For example, the confidence thresholds that determine an agency's preference for false positives or false negatives when using Amazon's Rekognition Web Service should be prominently exposed to staff and easily configurable. In these ways, contestable design expands front-line staff's knowledge of the policies and values embedded in the machine learning systems they use, while offering them opportunities to configure and interrogate them at run time. Such designs help agency staff learn about machine learning systems as the systems learn about agency staff. Machine learning systems, then, should be designed to allow users to both make key decisions about values-significant parameters and understand their significance. This requires moving away from defaults for these parameters and toward contestable systems that require engagement during setup and use.

Contestable design, moreover, is a prerequisite for continuous feedback from domain experts. Rather than traditional forms of contesting automated decisions—"out-of-band" processes (those external to the regular operation of the system itself, like exception handling and appeals)—contestable design brings argumentation, and therefore opportunities for learning and

293. Brauneis & Goodman, *supra* note 6, at 150.

recalibration within the system itself. Such continuous within-system learning is appropriate as "our models are, and will continue to be, fallible" and is particularly important in areas where the risks of " 'getting it wrong' can be quite high."[294] Active, critical, real-time engagement with the reasoning of machine learning systems' inputs, outputs, and models reduces the risk that machine learning systems will replace the logical and ethical frameworks that comprise expert agency judgment, and respond to risks posed by fallible models—regardless of whether fallibility is a product of design choices, shifts in policy, or inattention to gaps between phenomena and the representations we choose to capture them.

Contestability is a more active and dynamic principle than explanation[295] to guide design; for these reasons it breeds user engagement. Where the passivity of "explainable" algorithmic systems imagines engagement, reflection, and questioning as out-of-band activities—via exception handling, appeals processes, etc.—contestable systems foster active, critical engagement within the system. Explanations are also typically static, insofar as they are focused on conveying a single message; contestability, in contrast, aims to support interactive exploration of, and in some instances tinkering with, machine logic.

b) Design Should Trigger Human Engagement

Designing for contestability further requires the design of human-system interaction in a way that promotes an active and ongoing role for agency decision makers by overcoming the over-reliance on, and deference to, decision-support systems arising from automation bias—"the use of automation as a heuristic replacement for vigilant information seeking and processing"[296] and automation complacency—insufficient attention and monitoring of automation outputs.

It is certain that without thoughtful interventions, agency staff will be less attentive and engaged with the decisions supported by machine learning systems. Where the goal of introducing machine learning systems is to achieve a hybrid production of knowledge that builds on the strengths of human and machine ways of knowing—rather than to fully displace human decision making—chosen designs and policies must keep humans "in the game."

294. Hirsch et al., *supra* note 289, at 97.

295. *See id.* at 98; *see also* Daniel Kluttz et al., *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*, *in* AFTER THE DIGITAL TORNADO: NETWORKS, ALGORITHMS, HUMANITY (Kevin Werbach ed., forthcoming 2020).

296. Kathleen L. Mosier & Linda J. Skitka, *Automation Use and Automation Bias*, PROC. HUM. FACTORS & ERGONOMICS SOC'Y ANN. MEETING 344 (1999).

Research has established that particular design choices determine the extent to which expert judgment continues to be exercised in system-supported decision making in ways that skew policy regarding, for example, a preference for false-negatives over false-positives. The extent to which users substantively review a machine learning system output depends largely on whether the system signals that the result is anomalous or normal. For example, research has shown that radiologists scrutinize mammography films identified by decision-support systems as positive for cancer, catching many of the false positives that the system produced.[297] But films identified as normal rarely receive such inspection, allowing nearly all false negatives to evade detection. The perception that the system was over-inclusive—supported by the experience of identifying false positives—contributed to a belief in the infrequency of false negatives, when in fact the system systematically failed to identify certain images of cancer.

Similarly, in ongoing research with legal professionals using machine learning systems to aid in document review for discovery,[298] we have found that human review focuses on documents identified as relevant and thus appropriate for disclosure, according less attention to those the machine learning system designates as non-responsive. In both instances, perceptions of the machine's performance interact with experts' risk models to yield different levels of engagement with different outputs of the same underlying system. Yet, ideally we would want agency experts to be attentive to misidentifications or classifications (depending upon the task) produced by false positives and negatives—although, depending upon the use of the system, attention to one or the other might be directed by policy.

Iterative re-delegation of tasks and communicating about uncertainty are both design strategies that have been found successful as to maintain human skill and sense of responsibility and maintain attention to how machines are executing delegated tasks. For example, adaptive allocation,[299] in which an automated task is reallocated back to a human for periods of time, has been found to reduce automation complacency and improve subsequent attention

297. Anrey A. Povyakalo et al., *How to Discriminate Between Computer-Aided and Computer-Hindered Decisions: A Case Study in Mammography*, 33 MED. DECISION MAKING 98 (2013); *see* Goddard, *supra* note 182; *see also* Adrian Bussone et al., *The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems*, INT'L CONF. ON HEALTHCARE INFORMATICS 160 (2015) (discussing research findings on automation bias and self-reliance).

298. Daniel N. Kluttz & Deirdre K. Mulligan, *Automated Decision Support Technologies and the Legal Profession*, 34 BERKELEY TECH. L.J. 853 (2019).

299. *See* RAJA PARASURAMAN ET AL., THEORY AND DESIGN OF ADAPTIVE AUTOMATION IN AVIATION SYSTEMS, REP. NO. NAWCADWAR-92033-60 (July 17, 1992), https://apps.dtic.mil/dtic/tr/fulltext/u2/a254595.pdf [https://perma.cc/YBJ6-UAXQ].

to tasks by humans,[300] reduce human distraction, and promote a sense of responsibility.[301] Communicating the confidence that a system has in the conclusions it offered has been found to foster feelings of user responsibility[302] and, where coupled with a feedback loop, improve decisions in the moment and over time.[303]

c) Design Should Promote Contestation About Social and Political Values

Finally, for a number of reasons, contestability is of heightened importance where functions delegated to machine learning systems are deeply connected to social and political values such as fairness—values that are often ambiguous and contested.

First, given the numerous competing definitions of fairness,[304] there may well be multiple conflicting views on which definition should apply to a given function or in a given context; and when a definition of fairness is chosen, it may be susceptible to different formalizations.[305] We have more generally

300. *See* Raja Parasuraman, *Effects of Adaptive Task Allocation on Monitoring of Automated Systems*, 38 HUM. FACTORS 665 (1996).

301. *Id.*

302. Matthew Kay et al., *When (ish) Is My Bus?: User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems*, PROC. 2016 CONF. ON HUM. FACTORS COMPUTING SYSS. 5092 (2016) (finding that some bus riders felt that providing information about the uncertainty of an arrival time would make them more responsible for actions that led them to miss the bus: "you're more likely to be unhappy than if you missed the bus and can just blame the app").

303. Communicating the uncertainty related with the probabilistic nature of machine learning systems to users improves decision making, and if it is coupled with feedback mechanism, it can leverage human knowledge to increase accuracy of model over time. *See* Rocio Garcia-Retamero & Edward T. Cokely, *Communicating Health Risks with Visual Aids*, 22 CURRENT DIRECTIONS IN PSYCHOL. SCI. 392–93 (2013) (displaying a grid of pictograms, each representing a patient success or fatality improved the accuracy of people's risk assessment); *see also* Ulrich Hoffrage & Gerd Gigerenzer, *Using Natural Frequencies to Improve Diagnostic Inferences*, 73 ACAD. MED. 538 (1998) (noting that more medical experts could accurately estimate the positive predictive value (precision) of a test when presented with discrete counts or outcomes); Stumpf et al., *Toward Harnessing*, *supra* note 292, at 82 (noting that "user feedback has the potential to significantly improve machine learning systems").

304. *See* Deirdre K. Mulligan et al., *This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology*, PROC. 2019 ACM ON HUMAN-COMPUTER INTERACTION 3, 119 (2019) (defining the following terms: formal equality (blind to all other variables)—to each person an equal share; need-based—to each person according to individual need; effort-based—to each person according to individual effort; social contribution—to each person according to societal contribution; and merit-based—to each person according to merit).

305. For example, one could operationalize a given fairness definition around groups—seeing demographic parity or equal positive predictive values, or equal negative predictive values, or equal false positive or false negative rates, or accuracy equity—or one could

argued that, because of the strength and durability of design decisions, policymakers should be cautious about "baking" choices about human and public rights into technology systems and should steer such determinations to the least fixed point of technical intervention—permitting ongoing debate about such value choices and "designing technological hooks that permit different value choices in different contexts."[306] The discussion of the debates around what fairness required in COMPAS offers a clear example of the need to maintain visibility about these contested design elements.

Second, protecting and respecting values such as fairness and privacy may often hinge on *process*. Fairness and privacy, for example, are often *defined*—at least in part—by access to procedures that afford individuals meaningful participation, and to information (rules and data used to make decisions about them). These procedural aspects of values can be supported through contestable system design that minimizes the automation and opacity of those decisions and ensures that human judgment will be brought to bear in the ongoing shaping of, and in the assessment of the products of, decision-support systems.[307] Contestability keeps agency experts in control of these values questions, even while specific tasks and functions are handed off to machine learning systems. They can allow agency experts to revisit and tune machine-learning decisions to context-specific information that influences the perceived or actual fairness of a system. Contestable design responds to the demand of philosophers and ethicists that systems be designed to respond to diverse contexts ruled by different moral frameworks[308] and to support collaborative development of ethical requirements.[309]

V. CONCLUSION

In 1967 John Culkin, interpreting one of Marshal McLuhan's five postulates, offered the now-famous line: "We shape our tools and thereafter

---

operationalize it through an individual fairness metric, such as equal thresholds or devising a similarity metric. For a discussion, see *id.*

306. Mulligan & Bamberger, *supra* note 12, at 750.

307. *See* Min Kyung Lee & Su Baykal, *Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division*, PROC. 2017 ACM CONF. ON COMPUTER SUPPORTED COOPERATIVE WORK & SOC. COMPUTING 1035 (2017) (presenting discussion with others, and ability to interrogate systems logic can improve perceptions of fairness).

308. *See* Batya Friedman & Helen Nissenbaum, *Software Agents and User Autonomy*, PROC. 1ST INT'L CONF. ON AUTONOMOUS AGENTS 466 (1997).

309. *See* Matteo Turilli, *Ethical Protocols Design*, 9 ETHICS & INFO. TECH. 49 (2007).

they shape us."[310] The fear of abdicating important policy decisions to the control of tools, rather than autonomously wielding them in service of the public interest, lies at the heart of current concerns with the governmental adoption of machine learning systems. Such abdication strikes at the heart of administrative legitimacy and good governance, and suggests that machine learning systems are tools to be procured at our peril.

In many instances they are more texts than tools, and we suggest engaging them accordingly. They have their limits, their biases, and their blind spots. We should question and bicker with them, but we should also learn from and teach them. We should not blindly defer to them or bring them into our deliberations without knowing their backstories, working assumptions, and theories.

Through policy choices and design, we can build purposeful tools that are aligned with values chosen based on reason, expertise, transparency, and robust and ongoing deliberation and oversight. We can bind these new systems through carefully constructed "policy knots" that align them with the requirements of administrative law through policy, practice, and design.

310. John M. Culkin, *A Schoolman's Guide to Marshall McLuhan*, SATURDAY REV. 51, 70–72 (Mar. 18, 1967) (offering a "barously brief" distillation of Marshal McLuhan's writings, John M. Culkin expanded on one of McLuhan's five postulates, *Art Imitates Life*).

AUTOMATED DECISION SUPPORT TECHNOLOGIES AND THE LEGAL PROFESSION

Daniel N. Kluttz† *& Deirdre K. Mulligan*††

ABSTRACT

A quiet revolution is afoot in the field of law. Technical systems employing algorithms are shaping and displacing professional decision making, and they are disrupting and restructuring relationships between law firms, lawyers, and clients. Decision-support systems marketed to legal professionals to support e-discovery—generally referred to as "technology-assisted review" (TAR)—increasingly rely on "predictive coding": machine-learning techniques to classify and predict which of the voluminous electronic documents subject to litigation should be withheld or produced to the opposing side. These systems and the companies offering them are reshaping relationships between lawyers and clients, introducing new kinds of professionals into legal practice, altering the discovery process, and shaping how lawyers construct knowledge about their cases and professional obligations. In the midst of these shifting relationships—and the ways in which these systems are shaping the construction and presentation of knowledge—lawyers are grappling with their professional obligations, ethical duties, and what it means for the future of legal practice.

DOI: https://doi.org/10.15779/Z38154DP7K

† Postdoctoral Scholar, University of California, Berkeley School of Information, dkluttz@berkeley.edu.

†† Associate Professor, University of California, Berkeley School of Information, dmulligan@berkeley.edu. The authors thank the following for their helpful comments and feedback during the development of this paper: Kenneth Bamberger, Solon Barocas, Michael Buckland, Jenna Burrell, Tim Casey, Taylor Cruz, Madeleine Elish, Fernando Delgado, James X. Dempsey, Anne-Laure Fayard, Anna Lauren Hoffman, Karrie Karahalios, Nitin Kohli, Karen Levy, Clifford Lynch, Chris Mammen, Jonathan Marshall, Susan Nevelow Mart, Scott Skinner-Thompson, and Jennifer Urban. Special thanks to Ida Mahmoudi and Zane Kuser for their diligent and insightful research assistance, and Nitin Kohli for his expert analysis and feedback on this draft and collaboration on closely-related works. The authors also thank participants and audience members at the following workshops and conferences: the Berkeley Center for Law & Technology/Berkeley Technology Law Journal annual symposium, Data & Society Research Institute's "Algorithms on the Shop Floor" workshop, the Privacy Law Scholars Conference, the UC Berkeley School of Information's Friday Seminar series, the University of Colorado Law School colloquium, and WeRobot. Finally, we thank our interview subjects for their time and thoughtfulness. This research was supported in part by generous funding from the National Science Foundation (NSF INSPIRE: Value-Function Handoffs in Human-Machine Compositions, SES 1650589; NSF CCESTEM: Emerging Cultures of Data Science Ethics in the Academy and Industry), and the UC Berkeley Algorithmic Fairness and Opacity Working Group with financial support from Google.

Through in-depth, semi-structured interviews of experts in the e-discovery technology space—the technology company representatives who develop and sell such systems to law firms and the legal professionals who decide whether and how to use them in practice—we shed light on the organizational structures, professional rules and norms, and technical system properties that are shaping and being reshaped by predictive coding systems. Our findings show that AI-supported decision systems such as these are reconfiguring professional work practices. In particular, they highlight concerns about potential loss of professional agency and skill, limited understanding and thereby both over- and under-reliance on decision-support systems, and confusion about responsibility and accountability as new kinds of technical professionals and technologies are brought into legal practice. The introduction of predictive coding systems and the new professional and organizational arrangements they are ushering into legal practice compound general concerns over the opacity of technical systems with specific concerns about encroachments on the construction of expert knowledge, liability frameworks, and the potential (mis)alignment of machine reasoning with professional logic and ethics.

Based on our findings, we conclude that predictive coding tools—and likely other algorithmic systems lawyers use to construct knowledge and reason about legal practice—challenge the current model for evaluating whether and how tools are appropriate for legal practice. As tools become both more complex and more consequential, it is unreasonable to rely solely on legal professionals—judges, law firms, and lawyers—to determine which technologies are appropriate for use. The legal professionals we interviewed report relying on the evaluation and judgment of a range of new technical experts within law firms and, increasingly, third-party vendors and their technical experts. This system for choosing technical systems upon which lawyers rely to make professional decisions—e.g., whether documents are responsive, or whether the standard of proportionality has been met—is no longer sufficient. As the tools of medicine are reviewed by appropriate experts before they are put out for consideration and adoption by medical professionals, we argue that the legal profession must develop new processes for determining which algorithmic tools are fit to support lawyers' decision making. Relatedly, because predictive coding systems are used to produce lawyers' professional judgment, we argue they must be designed for *contestability*—providing greater transparency, interaction, and configurability around embedded choices to ensure decisions about how to embed core professional judgments, such as relevance and proportionality, remain salient and demand engagement from lawyers, not just their technical experts.

TABLE OF CONTENTS

I. INTRODUCTION

As applications based on advancements in fields such as cloud computing and machine learning have spread to the workplace, scholars and legal commentators have debated the extent to which such technical systems will affect markets for legal services, the practice of law, and the legal profession.[1] AI-based systems aimed at automating or assisting in lawyerly

1. *See, e.g.*, RICHARD SUSSKIND & DANIEL SUSSKIND, THE FUTURE OF THE PROFESSIONS: HOW TECHNOLOGY WILL TRANSFORM THE WORK OF HUMAN EXPERTS (2015); Daniel Martin Katz, *Quantitative Legal Prediction—Or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 EMORY L.J. 909 (2013); John O. McGinnis & Russell G. Pearce, *The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services*, 82 FORDHAM L. REV. 3041 (2014); Dana A. Remus & Frank Levy, *Can Robots Be Lawyers: Computers, Lawyers, and the*

tasks and decision making are currently being employed in a wide range of practice domains, including contract drafting and review, due diligence in mergers and acquisitions, risk-assessment in criminal justice settings, legal search and research, and document analysis and review in e-discovery, to name a few.[2]

Technology-assisted review (TAR), also called "predictive coding," systems for the discovery phase of litigation provide a particularly interesting example of a machine-learning-based ("ML-based") decision support system infiltrating a professional domain. Our research explores how professional identity, legal frameworks, interactions with clients and vendors, and organizational structures are shaping the adoption, use, and perceptions of TAR systems in the field of law.[3] Our interest in studying the adoption of machine learning tools in the legal profession was generated in part by a belief that lawyers—due to education and training, professional rules and ethical obligations, and their own interest in protecting themselves from professional liability—would place particularly stringent demands and expectations about the transparency, interpretability, configurability, and accountability of machine learning systems.

The concern that engineers and logics of automation will stealthily usurp or undermine the decision-making logics, values, and domain expertise of end-users has been an ongoing and legitimate complaint about decision-

---

Practice of Law, 30 GEO. J. LEGAL ETHICS 501 (2017); Tanina Rostain, *Robots versus Lawyers: A User-Centered Approach*, 30 GEO. J. LEGAL ETHICS 559 (2017); Sean Semmler & Zeeve Rose, *Artificial Intelligence: Application Today and Implications Tomorrow*, 16 DUKE L. & TECH. REV. 85 (2017); Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87 (2014); David C. Vladeck, *Machines without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117 (2014); John Markoff, *Armies of Expensive Lawyers, Replaced by Cheaper Software*, N.Y. TIMES (Mar. 4, 2011), https://www.nytimes.com/2011/03/05/science/05legal.html [https://perma.cc/K6NG-XEG3].

2. For overviews and discussions of such applications, see, for example, KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS: NEW TOOLS FOR LAW PRACTICE IN THE DIGITAL AGE (2017); Benjamin Alarie et al., *How Artificial Intelligence Will Affect the Practice of Law*, 68 U. TORONTO L. J. 106 (2018); Daniel Ben-Ari et al., *Artificial Intelligence in the Practice of Law: An Analysis and Proof of Concept Experiment*, 23 RICH. J.L. & TECH. 2 (2017); Kathryn D. Betts & Kyle Jaep, *The Dawn of Fully Automated Contract Drafting: Machine Learning Breathes New Life Into a Decades-Old Promise*, 15 DUKE L. & TECH. REV. 216 (2017); Richard Berk & Jordan Hyatt, *Machine Learning Forecasts of Risk to Inform Sentencing Decisions*, 27 FED. SENT'G REP. 222 (2015); Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in e-Discovery Can Be More Effective and More Efficient than Exhaustive Manual Review*, 17 RICH. J. L. & TECH. 1 (2011); David Lat, *How Artificial Intelligence Is Transforming Legal Research*, ABOVE L. (2018), https://abovethelaw.com/law2020/how-artificial-intelligence-is-transforming-legal-research/ [https://perma.cc/HD3J-YEGT] (last visited Aug. 25, 2019).

3. Our research is ongoing. We are continuing to interview lawyers, in-house technical professionals, and legal technology company representatives.

support and other computer systems.[4] As technology reconfigures work practices, researchers have documented potential loss of human agency and skill,[5] reduced opportunities to learn in the field,[6] both over- and under-reliance on decision-support systems,[7] confusion about responsibility,[8] and diminished[9] or exaggerated[10] accountability that leaves humans unable to exercise control but bearing the weight and blame for system failures.[11] For example, Elish explores how humans tend to take the brunt of failures in sociotechnical systems, acting as "moral crumple zones" by absorbing a

4. *See* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008) (identifying the slippage and displacement of case worker values by engineering rules embedded in an expert system); James H. Moor, *What is computer ethics?*, 16 METAPHILOSOPHY 266 (1985) (identifying three ways invisible values manifest in technical systems—hiding immoral behavior, gap-filling during engineering that invisibly embeds coders' value choices, and through complex calculations that defy values analysis); Jenna Burrell, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, 3 BIG DATA & SOC'Y 1 (2016) (describing three forms of opacity in corporate or state secrecy, technical illiteracy, and complexity and scale of machine-learning algorithms); Frank A. Pasquale, *Professional Judgment in an Era of Artificial Intelligence and Machine Learning*, 46 BOUNDARY 2, 73 (2019) (contrasting the reductionist epistemology and functionalist assumptions underlying substitutive automation with the holistic epistemology of professional judgment and the conflictual, political, and contestable nature of professional work, particularly in education and healthcare professionals).

5. *See* John D. Lee & Bobbie D. Seppelt, *Human Factors in Automation Design*, *in* SPRINGER HANDBOOK OF AUTOMATION, 417–36 (Shimon Y. Nof, ed., 2009) (detailing how automation that fails to attend to how it redefines and restructures tasks, and the behavioral, cognitive, and emotional responses of operators to these changes, produce various kinds of failure, including those that arise from deskilling due to reliance on automation).

6. *See* Matthew Beane, *Shadow learning: Building robotic surgical skill when approved means fail*, 64 ADMIN. SCI. Q. 87 (2019) (finding that robotic surgery limited the ability of medical residents to develop competence in traditional and approved ways so some residents resorted to "shadow learning" practices, which flaunted field norms and institutional policies, to gain surgical competence).

7. *See* Kate Goddard et al., *Automation bias: a systematic review of frequency, effect mediators, and mitigators*, 19 J. AM. MED. INFORMATICS ASS'N 121 (2011) (reviewing literature on automation bias in health care clinical decision support systems).

8. For an overview of research on technology-assisted decision making and responsibility, see Kathleen L. Mosier & Ute M. Fischer, *Judgment and Decision Making by Individuals and Teams: Issues, Models, and Applications*, 6 REVS. HUM. FACTORS & ERGONOMICS 198, 232–33 (2010).

9. *See* Judith Simon, *Distributed Epistemic Responsibility in a Hyperconnected Era*, *in* THE ONLINE MANIFESTO 145 (Luciano Floridi, ed., 2015); Helen Nissenbaum, *Computing and Accountability*, 37 COMMS. ACM 72 (1994).

10. Madeleine Clare Elish, *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*, 5 ENGAGING SCI., TECH., & SOC'Y 40, 40 (2019).

11. *See, e.g.*, Meg Leta Jones, *The ironies of automation law: tying policy knots with fair automation practices principles*, 18 VAND. J. ENT. & TECH. L. 77 (2015).

disproportionate amount of responsibility and liability relative to their actual control and agency.[12]

A. AUTOMATED DECISION MAKING IN THE LEGAL PROFESSION

Yet, there is little empirical research documenting how lawyers (as end-users) think about and incorporate ML-based decision support and knowledge discovery systems into their actual work practices. Studies of citation systems and software systems used to identify relevant case law have documented varied performance across technical systems on what lawyers would likely view as far less ambiguous tasks than defining and identifying documents for production (our focus in this Article). For example, analyses of citator accuracy have found wide discrepancies in performance.[13] Although ambiguous language in court opinions and different constructions of evaluative criteria—e.g., negative, distinguished, criticized—may cause some level of variation,[14] Hellyer argues that it does not fully explain the performance differences across citator services. He found the three leading citation services only agreed eleven percent of the time on the fact and type of negative treatment, and in a whopping eighty-five percent of the time, the three citators did not agree on whether there was negative treatment at all.[15] Hellyer concludes that while citators can be vastly improved, "users may need to reconsider the trust they place in citators" and "need to be aware of the citators' shortcomings."[16]

Research on knowledge discovery tasks in legal databases documents similar variation in performance. A recent study comparing the search results of six different legal database tools[17] found "hardly any overlap in the cases

12. Elish, *supra* note 10, at 6.

13. *See* Paul Hellyer, *Evaluating Shepard's, KeyCite, and BCite for Case Validation Accuracy*, 110 L. LIBR. J. 449, 450–55 (2018) (describing prior citator studies).

14. *Id.* at 454 ("[C]itation analysis is partly subjective. . . . [C]ourts sometimes use ambiguous language when discussing other cases . . . different researchers may have different ideas on what 'negative treatment' means, not to mention more specific terms like 'distinguished' or 'criticized.' ").

15. *Id.* at 464 ("[In] 357 citing relationships that have at least one negative label from a citator. . . . [A]ll three citators [BCite, KeyCite, and Shephard's] agree that there was negative treatment only 53 times. . . . [T]hat in 85% of these citing relationships, the three citators do not agree on whether there was negative treatment. . . . The three databases substantively agree on the type of negative treatment in only 40 of these citing relationships, which means that in this sample, they all agree with one another only 11% of the time.").

16. *Id.* at 476.

17. Susan Nevelow Mart, *The Algorithm as a Human Artifact: Implications for Legal [Re]Search*, 109 L. LIBR. J. 387, 390 (2017) (showing wide variation in results returned by six leading legal database providers for the same search terms and parameters). The six legal database providers studied were Casetext, Fastcase, Google Scholar, Lexis Advance, Ravel, and Westlaw.

that appear in the top ten results" with "[an] average of forty percent of the cases . . . unique to one database, and only about seven percent of the cases . . . returned . . . [by] all six databases."[18] Mart argues this evidences the "very different biases and assumptions" of product engineers, and concludes that "knowledge of this variability expands the opportunities for researchers to find relevant cases that can play 'some cognitive role in the structuring of a legal argument.' "[19]

Together, this research documents unexpectedly wide performance variations across software systems on knowledge discovery tasks that are essential to legal practice, and it highlights the need for "database providers to proactively think of algorithmic accountability as a way to improve research results for their users,"[20] as well as the need for lawyers to demand both more consistent performance across systems[21] and more information about the systems on which they rely for professional work.[22]

Despite their importance to today's practicing lawyer, there has been surprisingly little effort to examine how predictive technological systems are shaping legal practice and the field more broadly.[23] True, a few researchers have explored the performance of predictive coding e-discovery tools (our topic here), comparing them to each other and to the performance of human reviewers.[24] This research finds that technology-assisted review can

18. *Id.*

19. *Id.* at 390, 420 (quoting Stuart A. Sutton, *The Role of Attorney Mental Models of Law in Case Relevance Determinations: An Exploratory Analysis*, 45 J. AM. SOC'Y INFO. SCI. 186, 187 (1994)).

20. *Id.* at 420.

21. Hellyer, *supra* note 13 at 476 (arguing that "citing relationships are clear and can be objectively described," that "citators can and should do better," and that "citators can be reliable or they can be idiosyncratic, but they can't be both").

22. Mart, *supra* note 17, at 420; Hellyer, *supra* note 13 at 476 (arguing that while citators should be able to perform more consistently, "users may need to reconsider the trust they place in citators, and law librarians may need to rethink how they discuss citators with their patrons").

23. *See* Seth Katsuya Endo, *Discovery Hydraulics*, 52 U.C. DAVIS L. REV. 1317, 1337 (2019) ("[T]here is little empirical data about what drives lawyers' choices in their discovery practices.") (citing Judith A. McKenna & Elizabeth C. Wiggins, *Empirical Research on Civil Discovery*, 39 B.C. L. REV. 785, 803 (1998) ("Much of the literature on incentives affecting discovery practice is rooted in economic theory. Yet, there is little information about how lawyers actually make discovery decisions.")).

24. Maura R. Grossman & Gordon V. Cormack, *Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review*, 17 RICH. J. L. & TECH. 1 (2011); Herbert L. Roitblat et al., *Document categorization in legal electronic discovery: computer classification vs. manual review*, 61 J. AM. SOC'Y INFO. SCI. & TECH. 70 (2010); David Grossman, *Measuring and Validating the Effectiveness of Relativity Assisted Review*, EDRM WHITE PAPER SERIES (Feb. 2013) http://www.edrm.net/papers/measuring-and-validating-the-effectiveness-of-relativity

outperform human reviewers and save on attorney costs.[25] However, such studies have only taken place in controlled settings; they are not aimed at understanding how legal professionals work with these systems in practice.

The only empirical studies of decision support technologies and discovery practices at law firms that we have found were ethnographic studies undertaken in the 1990s by anthropologists and computer scientists tasked with designing systems to aid the litigation support team at a large law firm.[26] There, researchers found lawyers articulating and enacting a superior status relative to that of litigation support staff, whom they tended to see as only performing mundane, routine work of "document review" and incapable of the more complex decision making performed by attorneys.[27] Cautious and risk-averse, high in status, and protective of professional expertise, lawyers were therefore reluctant to hand off anything beyond what they saw as routinized work.

Finally, although not within the law firm setting, Christin's work on risk-recidivism algorithms offers important insights into how lawyers are interacting with a widely used set of algorithmic tools.[28] Her ethnographic study of algorithms in action in expert fields[29] reveals that professional

---

-assisted-review/ [https://perma.cc/VP9Y-BF8N] (reviewing the performance of one TAR product from a limited perspective).

25. *See, e.g.*, Grossman & Cormack, *supra* note 24, at 43, 48 (analyzing data collected during the course of the Text Retrieval Conference (TREC) 2009 Legal Track Interactive Task and finding that the predictive coding methods they reviewed "require, on average, human review of only 1.9% of the documents, a fifty-fold savings over exhaustive manual review" and concluding that TAR can yield more accurate results than exhaustive human review).

26. Jeanette Blomberg et al., *Reflections on a Work-Oriented Design Project*, 11 FOUND. & TRENDS HUM.-COMPUTER INTERACTION 237 (1996) (describing their experiences designing a case-based prototype system for information retrieval and their observations of organizational politics and divisions of labor between attorneys and litigation support staff at the law firm while designing image-analysis technologies to aid in document review and classification); Lucy Suchman, *Working Relations of Technology Production and Use*, 2 COMPUTER SUPPORTED COOPERATIVE WORK 21 (1993) (arguing for industrial designers to be aware of the work practices of not only technology production but also its use among various users, and describing observations of the status hierarchies, contestable knowledge claims, and actions of lawyers and litigation support staff at a law firm).

27. *See* Suchman, *supra* note 26, at 32 (describing how litigation support work was invisible to attorneys and how attorneys described such work as a "mindless, routine form of labor").

28. Angèle Christin, *Algorithms in Practice: Comparing Web Journalism and Criminal Justice*, 4 BIG DATA & SOC'Y 1 (2017).

29. *Id.* at 2. Christin distinguishes "expert fields" from professions (although they may overlap), defining expert fields as "configurations of actors and institutions sharing a belief in the legitimacy of specific forms of knowledge as a basis for intervention in public affairs." She makes this distinction for practical and strategic reasons. From a strategic standpoint, she makes the distinction in order to take a broader "field-based" analytical framework to

workers and managers appropriate machine-learning systems into their work practices as they do other technology: informed by routines, norms, obligations of professional identity, and their position relative to others within an organizational hierarchy (e.g., managers vs. workers).

Our review reveals notable gaps in the literature investigating the use of ML-based systems in the field of law. First, little is known about how legal professionals, their organizations, and their professional environments are shaping the adoption, implementation, and governance of machine-learning systems that support professional decision-making. This gap reflects the more general dearth of empirical data on professionals, their organizational environments, and their interactions with today's automated ML-based decision-making systems more generally. While research in computer science and the interdisciplinary FAT (Fairness, Accountability, and Transparency) community interrogates and evaluates the technical workings of such systems to shed light on values and ethics,[30] and an increasing amount of legal scholarship theorizes and makes normative claims about what laws or regulatory frameworks we need to address automated decision making systems,[31] there is little rigorous, empirical social science research into the

her sociological analysis (drawing specifically on Pierre Bourdieu's conception of a "field"). We take an analogous, if conceptually distinct, systems-based view of the legal profession and lawyers, in which the profession is marked by constantly evolving processes of conflict, cooperation, and exchange with internal and external stakeholders. *See* ANDREW ABBOTT, THE SYSTEM OF PROFESSIONS: AN ESSAY ON THE DIVISION OF EXPERT LABOR (1988). And from a practical standpoint, she makes the distinction between expert fields and profession so that she can include in her comparative analysis journalists, who are not typically thought of as a highly professionalized occupation in the sense that, compared to highly professionalized domains like law, they lack, among other things, state-licensed monopoly control over the barriers to entry for their work. By staying within the legal profession to understand how lawyers are appropriating machine-learning decision support systems, we make no comparisons across professions here and thus see no need to follow Christin's distinction between "expert fields" and professions. For more on Bourdieu's vision of "fields," see PIERRE BOURDIEU & LOÏC J.D. WACQUANT, AN INVITATION TO REFLEXIVE SOCIOLOGY (1992). For more on sociological field theory more generally, see Daniel N. Kluttz & Neil Fligstein, *Varieties of Sociological Field Theory*, *in* HANDBOOK OF CONTEMPORARY SOCIOLOGICAL THEORY 185 (Seth Abrutyn ed., 2016).

30. *See, e.g.*, Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153 (2017); Amit Datta et al., *Discrimination in Online Advertising: A Multidisciplinary Inquiry*, 81 PROC. MACHINE LEARNING RES. 20 (2018); Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, 8TH INNOVATIONS IN THEORETICAL COMPUT. SCI. CONF. 43 (2017); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017).

31. *See, e.g.*, RYAN CALO ET AL., ROBOT LAW (2016); Danielle Keats Citron & Frank A. Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83 (2017); We Robot Conference (2019), https://robots.law.miami.edu [https://perma.cc/UC8F-B8E6] (last visited Sept. 20, 2019) (providing background papers and presentations).

professionals, their organizations, and the broader professional ecosystems in which these technical systems are embedded.[32] How are these systems entering the field of law? What kinds of socio-material forces—for example, professional ethical duties, identity as an expert, the configuration of the system itself—affect how professionals understand and use such systems? How is the introduction of automated decision-making systems shaping professional practices and the profession?

Second, little is known about how professionals are interacting with the new wave of predictive algorithmic systems—particularly ML-based systems—that are being sold as aids for professional decision making. While traditional, rule-based expert systems have had a long history in professional fields like medicine, there are unique challenges posed by today's predictive algorithmic systems. Whereas engineers of expert systems explicitly program in a set of rules—ideally based on the domain knowledge of adept subject matter experts—today's ML-based systems are designed, in effect, by deriving a set of decision rules from the data on which they train, which creates some unique challenges to ensuring systems accord with professional expertise and judgment. Some of the algorithms used by such systems make it difficult to understand the rules they have learned from the data. Unlike an expert system where domain professionals can review and interrogate rules, these systems can provide insight into the inputs and outputs but lack the ability to easily interrogate the rules or the reasoning by which the outputs were generated. In addition, today's predictive ML-based systems are dynamic, usually probabilistic, and therefore do not have one "right" answer. This plasticity challenges even the limited oversight provided by examinations of inputs and outputs. The consequence of such characteristics is that predictive algorithmic systems embed many subjective judgments on the part of system designers—for example, judgments about training data, how to clean the data, how to weight different features, which algorithms to use, what information to emphasize or deemphasize, etc. We know little about whether and how the distinct features of these new tools affect professionals' interactions with them.

Our empirical research begins to fill this gap. We focus on AI-based e-discovery systems and how lawyers go about conducting discovery in today's world of electronically stored information (ESI). Specifically, we study lawyers' use of TAR, also called "predictive coding," systems. TAR is used to identify documents relevant, responsive, and not protected by legal

32. Admittedly, this may, in part, be an artifact of researchers tending to focus on domains such as "predictive policing," risk-recidivism, facial recognition, ad targeting, or recommender systems, where professionals may play a limited role in the adoption, use, and governance of the systems.

privileges, to the opposing party's discovery requests. Our analysis of the relations among lawyers, litigation support professionals (whether inside or outside the firm), and predictive coding technologies in the modern-day legal services field provides a rich account of the actual effect of machine learning systems on legal practice and identifies domain-specific challenges posed by current ML-based systems and practices. The insights we offer raise new questions for the profession, and we identify new sites for interventions to shape the adoption, use, and governance of these tools going forward.

II. TAR AND PREDICTIVE CODING FOR E-DISCOVERY

One of the main challenges facing litigants today is the time and expense required to wade through ever-increasing amounts of ESI during discovery (e.g., data produced by smartphones, email, wearable devices, and the Internet of Things). Lawyers must review their clients' records in order to search, collect, and produce those that are relevant and responsive to the other party's requests and not protected by legal privileges. Discovery was an onerous process even in the days prior to ESI. With the vast amounts of ESI today, however, e-discovery can take inordinate amounts of time for lawyers and clients tasked with manually reviewing ESI. It also entails huge monetary costs for litigants. A 2012 study by RAND researchers found that e-discovery production costs averaged about $18,000 per gigabyte of information, with costs attributable to document review being seventy percent or more of total e-discovery costs in more than half of the fifty-seven cases studied.[33] The stakes of e-discovery are also high in other ways, with attorneys and clients exposed not only to the risk of adverse case outcomes but also to potential loss of attorney-client or other confidentiality privileges—and even disciplinary action—if they inadvertently produce non-discoverable ESI or withhold discoverable ESI .

The concomitant rise of ESI and advancements in technology over the past two decades or so have spawned an ever-growing industry of e-discovery specialists, support staff, consultants, technology vendors, and products. Predictive coding systems, under the umbrella of TAR, are marketed as tools to aid legal professionals in managing, classifying, and reviewing ESI at a fraction of the cost of traditional manual review.[34]

33. *See* NICHOLAS M. PACE & LAURA ZAKARAS, WHERE THE MONEY GOES: UNDERSTANDING LITIGANT EXPENDITURES FOR PRODUCING ELECTRONIC DISCOVERY (2012) (analyzing the collection, processing, and review costs for e-discovery across fifty-seven cases).

34. Examples of TAR products and e-discovery platforms on the market today include those from Brainspace (https://www.brainspace.com/ [https://perma.cc/RRQ9-XLZ6]), Catalyst (https://catalystsecure.com/ [https://perma.cc/XN6J-E469]), Exterro (https://www.exterro.com/e-discovery-software/data-management/predictive-intelligence/

Research supports the assertion that predictive coding can save on attorney review time, and thus costs.[35] Our research focuses on these e-discovery systems because, based on our conversations with lawyers and review of the literature, they represent one of the most well-developed applications of automated decision-support technology in the legal profession to date and because they provide a useful lens through which to discuss particular professional and ethical issues in their design and use.[36]

Broadly, TAR encompasses a number of technologies and techniques used on ESI, such as machine learning, clustering, semantic analysis, and sentiment analysis, to accomplish a broad range of tasks (e.g., email threading, de-duplication, document classification, visualization) that may or may not use predictive algorithms or machine-learning techniques to predict potentially responsive documents. Although most in the industry use "TAR" and "predictive coding" interchangeably, for simplicity, and because we want to focus attention on the ML-based process of analyzing and predicting which documents among a corpus are responsive and not responsive during discovery, we follow industry convention and use "predictive coding" unless TAR is specifically used in quotes from the literature or our interviews.[37]

There are different machine-learning techniques used in predictive coding, requiring varying levels of human reviewer effort and varying degrees

---

[https://perma.cc/C7AC-RNCG]), H5 (https://www.h5.com/ [https://perma.cc/UE45-9E4Q]), Nuix-Ringtail (https://get.nuix.com/nuix-ringtail/ [https://perma.cc/R3GB-ARXX]), and Relativity (https://www.relativity.com/ [https://perma.cc/BT3E-BCXT]).

35. *See, e.g.*, Grossman & Cormack, *supra* note 24, at 43 (analyzing data collected during the course of the Text Retrieval Conference (TREC) 2009 Legal Track Interactive Task and finding that the TAR methods they assessed "require, on average, human review of only 1.9% of the documents, a fifty-fold savings over exhaustive manual review").

36. Katie Shilton, *Values and Ethics in Human-Computer Interaction*, 12 FOUND. & TRENDS HUM.-COMPUTER INTERACTION 107 (2018).

37. See, for example, definitions of "predictive coding" provided by the Electronic Discovery Reference Model (EDRM), a community of e-discovery and legal professionals housed at Duke University Law School EDRM. *Predictive Coding*, EDRM GLOSSARY, https://www.edrm.net/glossary/predictive-coding/ [https://perma.cc/6RYE-PNH8] (defining "predictive coding" as a subset of TAR tools that incorporate machine learning to distinguish relevant from non-relevant documents). Despite sharing some underlying general principles, predictive coding as used in the specific context of legal discovery should not be confused with predictive coding concepts and models as developed in neuroscience, cognitive science, and machine learning. For a review of predictive coding in these fields, see Yanping Huang & Rajesh P. N. Rao, *Predictive Coding*, 2 WILEY INTERDISC. REVS. 580 (2011); Geoffrey E. Hinton, *Learning Multiple Layers of Representation*, 11 TRENDS COGNITIVE SCI. 428 (2007); and Andy Clark, *Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science*, 36 BEHAV. & BRAIN SCI. 181 (2013). For applications to signal processing and data compression, see YUN Q. SHI AND HUIFANG SUN, IMAGE & VIDEO COMPRESSION FOR MULTIMEDIA ENGINEERING: FUNDAMENTALS, ALGORITHMS, AND STANDARDS (1999).

of initial training.[38] Cormack and Grossman, two leading experts on predictive coding, currently argue for continuous active learning (CAL), in which a subject-matter expert (in our case, an attorney) can continue to adjust the training algorithm during document review, as the best available method. However, as we will discuss, our interviews of attorneys and e-discovery suggest that, in practice, there is disagreement and confusion over defining, measuring, and achieving optimal precision, recall, and human reviewer effort.[39]

A. GOVERNANCE FRAMEWORK FOR ETHICAL AND RESPONSIBLE USE OF PREDICTIVE CODING

The governance of the legal profession's use of predictive coding is based in normative principles of responsible conduct during discovery and professional ethical duties governing lawyers. Thus, regulation of predictive coding for e-discovery emanates from general principles and guidelines. To the extent that such principles are formalized, they are found in jurisdiction-specific rules of civil procedure, case law, and, most importantly for our purposes, ethical rules of professional conduct as defined by state bars.[40]

The Federal Rules of Civil Procedure (FRCP), while not speaking directly to predictive coding, were amended in 2015 to give new guidance regarding e-discovery.[41] Overall, the amended rules reflected courts' desire to

38. *See* Cormack et al., *Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery*, PROC. 37TH INT'L ACM SIGIR CONF. ON RES. & DEV. INFO. RETRIEVAL, 153 (2014); *see also* Grossman et al., *Automatic and Semi-Automatic Document Selection for Technology-Assisted Review*, PROC. 40TH INT'L ACM SIGIR CONF. ON RES. & DEV. INFO. RETRIEVAL, 905 (2017).

39. *See* Cormack et al., *supra* note 38, at 161 (comparing Continuous Active Learning (CAL), Simple Passive Learning (SPL), and Simple Active Learning (SAL) and concluding that CAL exhibited fewer limitations and superior performance to the other predictive coding approaches).

40. We focus on formal governance mechanisms here—e.g., rules of court, judicial opinions, and ethical duties instantiated in professional rules, all backed by the sanctioning power of regulatory bodies or courts.

41. In the interest of space, we do not discuss all of the 2015 FRCP amendments as they pertain to e-discovery. For summaries, see *FRCP & E-Discovery: The Layman's Guide*, EXTERRO (2017), https://www.exterro.com/resources/frcp-e-discovery-pdf-guide/ [https://perma.cc/2ZZ3-GW9R]; *see also* Stephanie Serhan, *Calling an End to Culling: Predictive Coding and the New Federal Rules of Civil Procedure*, 23 RICH. J. L. & TECH. 1 (2017) (reviewing the 2015 amendments to the FRCP as applied to predictive coding, the split at that time among courts over when to use predictive coding during a case, and arguing that predictive coding should be done at the outset of discovery on the entire set of ESI rather than an already-keyword-culled set of documents); *see also* Karl Schieneman & Thomas C. Gricks III, *The Implications of Rule 26(g) on the Use of Technology-Assisted Review*, 7 FED. CTS. L. REV. 239, 247–84 (2013) (evaluating the proper way to use TAR at different phases of e-discovery for

encourage cooperation and accountability among parties, to promote speedy and efficient litigation, and, with amendments to Rule 26, to emphasize the principle of proportionality when it comes to the scope of discovery.[42] Regarding proportionality, the proportionality factors previously found in Rule 26(b)(2)(C)(iii) were amended and made more explicit by moving them to Rule 26(b)(1), which now reads as follows:

> Scope in General. Unless otherwise limited by court order, the scope of discovery is as follows: Parties may obtain discovery regarding any non-privileged matter that is relevant to any party's claim or defense and proportional to the needs of the case, considering the importance of the issues at stake in the action, the amount in controversy, the parties' relative access to relevant information, the parties' resources, the importance of the discovery in resolving the issues, and whether the burden or expense of the proposed discovery outweighs its likely benefit. Information within this scope of discovery need not be admissible in evidence to be discoverable.[43]

As to predictive coding specifically, although it is usually not required in a case and instead agreed to by the parties in a case, courts and the legal profession more generally have taken notice of its rise and have, on the whole, welcomed its use, although some legal commentators have raised concerns. For example, Endo argues that the opacity of ML-based predictive coding systems can undermine the due process norm of participation, especially for parties who lack adequate understanding of the system's reasoning process.[44] Remus recognizes predictive coding's potential benefits but cautions that it also brings significant costs: 1) the tendency of attorneys and judges to overlook the wide variation in predictive coding systems' technical features and efficacy, 2) the erosion of lawyers' professional jurisdiction over discovery by lowering professional oversight standards and by delegating the process to non-lawyer computing systems, vendors, and technical specialists, and 3) the undermining of client representation, with threats to work-product and attorney-client privileges and confidentiality via new rules and norms pushing lawyers to cooperate with the opposing party by disclosing things like seed sets or system-evaluation metrics.[45]

---

purposes of the attorney's reasonable-inquiry and certification requirements under Rule 26(g)).

42. *See FRCP & E-Discovery: The Layman's Guide*, *supra* note 41, at 5.

43. Fed. R. Civ. P. 26(b)(1) (emphasis added).

44. Seth Katsuya Endo, *Technological Opacity & Procedural Injustice*, 59 B.C. L. REV. 821 (2018).

45. Dana A. Remus, *The Uncertain Promise of Predictive Coding*, 99 IOWA L. REV. 1691 (2014).

Despite these cautions, judges have begun to approve the use of predictive coding in published opinions. For example, in *Da Silva Moore v. Publicis Groupe*, an employment discrimination case involving a high volume of electronic documents (over three million emails), U.S. Magistrate Judge Andrew Peck became the first federal judge to publish an opinion explicitly approving the use of computer-assisted review software as an acceptable of means of conducting e-discovery in appropriate cases.[46] There, the parties had agreed to use predictive coding prior to discovery, but the plaintiffs disputed the scope and defendants' implementation of it as detailed in the e-discovery protocol. While Judge Peck condoned predictive coding in "appropriate" circumstances, he did not specify fixed requirements of appropriateness and instead looked to the facts of the case in their entirety.[47] Among other relevant facts, predictive coding was cost-effective in this case compared to manual review given the large volume of documents, the parties had agreed to its use at the outset, defense counsel had been transparent in sharing its procedures with plaintiffs in its e-discovery protocol (e.g., disclosing the seed set used to train the predictive coding system), and counsel otherwise complied with the FRCP governing discovery. In particular, Judge Peck emphasized the importance of defense counsel satisfying the proportionality requirements of Rule 26 of the FRCP.[48] Although beyond our scope here, subsequent cases have followed with similar reasoning,[49] and professional bodies have convened to address the evolving landscape of e-discovery technologies and issue best practices.[50]

---

46. Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 193 (S.D.N.Y. 2012) ("This Opinion appears to be the first in which a Court has approved of the use of computer-assisted review. That does not mean computer-assisted review must be used in all cases, or that the exact ESI protocol approved here will be appropriate in all future cases that utilize computer-assisted review. . . . What the Bar should take away from this Opinion is that computer-assisted review is an available tool and should be seriously considered for use in large-data-volume cases where it may save the producing party (or both parties) significant amounts of legal fees in document review.").

47. *Id.* (emphasis added) ("As with keywords or any other technological solution to e-discovery, counsel must design an *appropriate* process, including use of available technology, with *appropriate* quality control testing, to review and produce relevant ESI while adhering to Rule 1 and Rule 26(b)(2)(C) proportionality.").

48. Again, prior to the 2015 amendments, e-discovery proportionality factors were previously found at Rule 26(b)(2)(C)(iii) of the FRCP.

49. *See, e.g.*, Nat'l Day Laborer Org. Network v. U.S. Immigration & Customs Enf't Agency, 877 F. Supp. 2d 87, 109 (S.D.N.Y. 2012) ("[P]arties can (and frequently should) rely on . . . machine learning tools to find responsive documents."); Dynamo Holdings Ltd. P'ship v. Comm'r, 143 T.C. 183, 191–92 (2014) ("Although predictive coding is a relatively new technique, and a technique that has yet to be sanctioned (let alone mentioned) by this Court in a published Opinion, the understanding of e-discovery and electronic media has advanced significantly in the last few years. . . . In fact, we understand that the technology

Finally, with respect to professional ethical rules and technology, the duty of attorney competence is most applicable to lawyers' use of ML-based predictive coding. The American Bar Association's (ABA) Model Rule 1.1 of the Model Rules of Professional Conduct states: "A lawyer shall provide competent representation to a client. Competent representation requires the legal knowledge, skill, thoroughness and preparation reasonably necessary for representation."[51] In 2012, the legal profession began the process of establishing a legal duty of *technological* competence on lawyers when the ABA's House of Delegates amended Comment 8 to Model Rule 1.1 to read: "To maintain the requisite knowledge and skill, a lawyer should keep abreast of changes in the law and its practice, including the benefits and risks associated with relevant technology, engage in continuing study and education and comply with all continuing legal education requirements to which the lawyer is subject."[52]

As of the end of February 2019, thirty-six states have formally adopted the amended comment to Rule 1.1.[53] On February 26, 2019, Texas became the most recent state to adopt the ABA's Comment 8 to Rule 1.1, when the Supreme Court of Texas amended Paragraph 8 of the comment to Rule 1.01 of the Texas Disciplinary Rules of Professional Conduct comment to track the ABA's model language.[54]

Although California has not specifically adopted the language of the ABA's Comment 8 to Rule 1.1 into its own rule of professional conduct

---

industry now considers predictive coding to be widely accepted for limiting e-discovery to relevant documents and effecting discovery of ESI without an undue burden."); Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125, 126 (S.D.N.Y. 2015) (holding that TAR is "an acceptable way to search for relevant ESI in appropriate cases").

50. *See, e.g.*, *The Sedona Conference, The Sedona Principles, Third Edition: Best Practices, Recommendations & Principles for Addressing Electronic Document Production*, 19 SEDONA CONF. J. (2018); *Frameworks and Standards: Technology Assisted Review*, EDRM, (2018), http://www.edrm.net/frameworks-and-standards/technology-assisted-review/ [https://perma.cc/932Y-BH8S] (last visited Sept. 20, 2019).

51. MODEL RULES OF PROF'L CONDUCT r. 1.1 (AM. BAR ASS'N 2017).

52. *Id.* at cmt. 8 (reviewing and explaining relevant legal standards and ethical duties regarding attorney technological competence in e-discovery).

53. Robert Ambrogi, *36 States Have Adopted Ethical Duty of Technology Competence*, ROBERT AMBROGI'S LAWSITES (Mar. 12, 2019), https://www.lawsitesblog.com/tech-competence/ [https://perma.cc/ZG28-KHAR] (providing running tally of states that have adopted the ABA's comment to Model Rule 1.1 and links to each state's rule).

54. *See* Order Amending Comment to the Texas Disciplinary Rules of Professional Conduct, Misc. Docket No. 19-9016 (Tex. Feb. 26, 2019), http://www.txcourts.gov/media/1443638/199016.pdf [https://perma.cc/EP3S-YQQD] ("Because of the vital role of lawyers in the legal process, each lawyer should strive to become and remain proficient and competent in the practice of law, *including the benefits and risks associated with relevant technology*.") (emphasis added).

regarding competency,[55] the State Bar of California has, since 2015, nevertheless incorporated the model rule's duty of technology competence with respect to e-discovery via a formal ethics opinion.[56] This opinion is particularly instructive not only because California is home to a thriving technology sector but also because it provides an extended discussion of attorney competence specifically as applied to conducting e-discovery during litigation. Attorneys in California should be able to perform the following nine skills:

> 1) initially assess e-discovery needs and issues, if any;
>
> 2) implement/cause to implement appropriate ESI preservation procedures;
>
> 3) analyze and understand a client's ESI systems and storage;
>
> 4) advise the client on available options for collection and preservation of ESI;
>
> 5) identify custodians of potentially relevant ESI;
>
> 6) engage in competent and meaningful meet and confer with opposing counsel concerning an e-discovery plan;
>
> 7) perform data searches;
>
> 8) collect responsive ESI in a manner that preserves the integrity of that ESI; and

---

55. California's professional rule regarding attorney competence is Rule 3-110 of the Rules of Professional Conduct of the State Bar of California. It holds:

> (A) A member shall not intentionally, recklessly, or repeatedly fail to perform legal services with competence.
>
> (B) For purposes of this rule, "competence" in any legal service shall mean to apply the 1) diligence, 2) learning and skill, and 3) mental, emotional, and physical ability reasonably necessary for the performance of such service.
>
> (C) If a member does not have sufficient learning and skill when the legal service is undertaken, the member may nonetheless perform such services competently by 1) associating with or, where appropriate, professionally consulting another lawyer reasonably believed to be competent, or 2) by acquiring sufficient learning and skill before performance is required.

RULES OF PROF'L CONDUCT, r. 3-110 (CAL. ST. BAR ASS'N 1992).

56. Cal. St. Bar Standing Comm. on Prof'l Responsibility & Conduct, Formal Op. No. 2015-193 at 3 (2015) (quoting the revised Comment 8 to ABA Model Rule 1.1 to state that "[m]aintaining learning and skill consistent with an attorney's duty of competence includes keeping 'abreast of changes in the law and its practice, including the benefits and risks associated with relevant technology . . . ' ").

> 9) produce responsive non-privileged ESI in a recognized and appropriate manner.[57]

In California, as in other states adopting the revised comment to the ABA model rule, if an attorney does not possess the requisite skills described above, they can satisfy their ethical obligation of e-discovery technology by associating with competent co-counsel or expert consultants. Such an expert could be a vendor, a subordinate attorney, or even the client itself, as long as they possess the necessary expertise.[58] However, associating with an expert raises another ethical duty—the duty to supervise—and potential tensions regarding professional expertise between attorneys, technology vendors, and clients that we address in our findings. Even if an attorney associates with a co-counsel or e-discovery consultant with expertise in handling e-discovery technology, that attorney still has the responsibility to supervise such an expert and is ultimately responsible for the work of the expert.[59]

Finally, in its ethics opinion, the California State Bar Standing Committee explicitly did not define a standard of care for attorneys for liability purposes, and it reserves disciplinary action for situations where a lawyer intentionally, recklessly, or repeatedly demonstrates a lack of competence. Our review indicates that subsequent situations involving ethical duties of competence and predictive coding have not been formally adjudicated before such judicial or disciplinary bodies, so it remains an open question as to just how accountable such bodies will hold lawyers when it comes to professional ethics. Given that AI-based products—for e-discovery as well as other legal tasks—will only grow in application and reach, we need further research into how lawyers actually understand and use them in practice.

III. RESEARCH DESIGN

We draw primarily on qualitative evidence obtained from approximately twenty-six hours of semi-structured, in-depth interviews of twenty-five respondents who work with predictive systems in the legal profession—attorneys, litigation support staff working in law firms, and managers at

57. *Id.* at 3–4.

58. *Id.* at 5.

59. *Id.* The rule governing attorney competence in California, cited within the 2015 ethics opinion as Rule 3-110 of the Rules of Professional Conduct of the State Bar of California, is now cited as CA ST RPC Rule 1.1 (Business and Professions Code Section 6068(e)) (new rules approved by the Supreme Court of California May 10, 2018, effective Nov. 1, 2018).

companies that provide decision-support technology products and services to lawyers.[60]

Of our twenty-five respondents, seventeen work at law firms (twelve attorneys, five litigation/technical support staff at law firms), and eight work at legal technology companies. Within this latter group, all of their positions are at the management level: CEO (two), CTO (two), COO (one), Vice President (one), Director of Consulting (one), and Litigation Manager (one).[61] All respondents are based in the United States, although their firms/organizations do business overseas, as well.

Our sample of attorneys is not representative of the population of attorneys in the United States, of course. For example, attorneys in our sample all work at law firms with greater than fifty attorneys and, with the exception of one respondent attorney working at a large plaintiff-oriented firm, would be classified roughly as corporate defense law firms. Because we focus on decision-support tools applied to the e-discovery context, all law firms represented have significant litigation practices. Our focus on attorneys and legal tech managers working with these kinds of law firms was strategic, as our research indicates that they are the firms most likely to be targeted as potential customers by technology company vendors and are the firms most likely to have the clients, resources, and types of cases (i.e., cases with significantly large volumes of ESI) that call for the use of AI-based systems for e-discovery, such as predictive coding tools.[62] In other words, they are the

60. This research is ongoing, so numbers and findings may change. We conducted interviews in-person or, if respondents were not available to interview in-person, over the phone. All in-person interviews took place in the Bay Area of California. Interview procedures were approved by and complied with the University of California, Berkeley's Office for the Protection of Human Subjects.

61. Three of these respondents are also founders of their companies (2 Founder/CEO; 1 Founder/CTO).

62. *See* Drew Simshaw, *Ethical Issues in Robo-Lawyering: The Need for Guidance on Developing and Using Artificial Intelligence in the Practice of Law*, 70 HASTINGS L.J. 173, 193 (2019) (noting that the rise of AI in e-discovery could inhibit access to justice because "the benefits of AI-driven e-discovery might, at least at first, only be recognized by large firms because many smaller practices lack designated e-discovery units"); *see also* Sean Semmler & Zeeve Rose, *Artificial Intelligence: Application Today and Implications Tomorrow*, 16 DUKE L. & TECH. REV. 85, 90 (2017). Semmler and Rose state:

> [There is a] possibility that big firms, with their resources and profit margins, are well situated to gain access to this disruptive technology at an earlier stage than smaller firms. Subscriptions to legal A.I. applications may be expensive (early on), and if big firms can buy this technology, become familiar with it now, and use it to attract new clients while retaining their old clientele, then by the time smaller firms get access to the same technology, it may be too late.

Id.

firms and attorneys most likely to have experience with and knowledge of these technologies. Thus, to the extent that our data identify challenges posed by the introduction and use of such systems, conclusions we draw are likely to be conservative, if anything.

IV. THE RISE OF TAR AND PREDICTIVE CODING IN THE LEGAL PROFESSION

To what do lawyers and legal professionals in the surrounding legal services environment attribute the rise of predictive coding systems, and TAR more generally, in the legal profession?

A. COST-CUTTING

Our respondents consistently positioned TAR as a cost-cutting strategy. Like the well-established practice of outsourcing to contract attorneys, and to out-of-country attorneys, delegating "document review" and e-discovery tasks to technical tools is viewed as a way to reduce litigation costs. The rising costs of litigation are a product of both escalating lawyer fees and the explosion in electronic documents produced by daily corporate activities in the digital age. Our respondents viewed TAR and predictive coding primarily as a response to this unprecedented growth in ESI. As Carrie Lewis, a partner at a law firm representing corporate clients, explained:

> We're seeing more and more that the general counsel has to show to their leadership and to their board that they have reduced costs by X percent or increased the use of technology. Then they're coming to us and saying how do we measure this? How do we show this? What do we do?[63]

Jason Ellison, currently a manager at a vendor and formerly a litigation support specialist within a law firm, echoed that sentiment: "Clients are increasingly looking at their spends. They're increasingly analyzing line items on bills and pushing down on law firm clients. This is something that started to get a lot of attention about ten years ago."[64] And Samir Anand, an attorney respondent, spoke to the issue of lawyer fees: "What's happened is lawyer rates have gone up so high that everyone just assumes that it's [predictive coding] being done by somebody else."[65]

The increased reliance on technology to facilitate document review, combined with the growing practice of procuring legal services, has fueled

63. Interview with Carrie Lewis, attorney (Jan. 10, 2019).

64. Interview with Jason Ellison, manager at a vendor and former litigation support specialist (Jan. 17, 2019).

65. Interview with Samir Anand, attorney (Jan. 4, 2019).

the growth of legal support vendors that manage and secure the voluminous corpus of documents generated by a business' general operation and provide document review services.[66] Our interviewees explained that the evolution of TAR tools and pressure to cut costs is reorganizing the relationship between lawyers, clients, and these vendors. This reorganization takes several forms. As to payment, most commonly, a corporate client (i.e., a party to the litigation) pays the e-discovery vendor, but the day-to-day interactions take place between the vendor and the attorneys. As Ken Summers, an executive at a TAR vendor, explained: "The vast majority of our clients are corporations, typically in the Fortune 200[.] In 90%-plus of the cases, [they] pay us. They're the true client. The client we work with on a day-to-day basis is the law firm that represents those corporate clients. That's the structure."[67] Increasingly, our respondents indicate, legal tech vendors are also marketing their services not only to law firms but directly to the corporations. One respondent, who wished not to be recorded but had managed an overseas office of e-discovery technicians and document reviewers for an e-discovery vendor, described her experience of being told by her supervisor to skip going through the law firm with which her company had contracted and go straight to the corporate client to attempt to cultivate a direct relationship with the corporation in the hopes of future work.

The pressure to reduce legal costs, combined with the increase of technically sophisticated vendors offering complex computational systems to aid legal decision making, has also led to reconfigurations in the extent to which law firms have control over the technologies that they use for their own work. Paul Young, an executive at an e-discovery vendor, explained:

> The reality is if you're a multi-billion dollar company, do you really want a law firm that charges $1,000 an hour making all of your decisions for you? Or do you want to have people internally [who are] definitely looking out for your best interests and vetting outsourced vendors accordingly, contracting directly with them, managing that process internally versus going through a law firm?[68]

This seems to be the case particularly with larger corporate clients. As one partner at a law firm told us:

> The overwhelming trend is that the lawyers are being taken out of that process . . . decisions about which corporate lawyers to use have been centralized by the client—which time-entry programs to use, which billing software to use, which ways that we report to the

66. Silvia Hodges Silverstein, *What We Know and Need to Know about Legal Procurement*, 67 S.C. L. REV. 485 (2016).

67. Interview with Ken Summers, executive at a TAR vendor (Feb. 04, 2019).

68. Interview with Paul Young, e-discovery vendor executive (Jan. 16, 2019).

> client—are all governed by terms and conditions from the client at the beginning of the relationship . . . .[69]

Angus Martin, a partner at a law firm, focused instead on the preferred-provider aspect of his dealings with large corporate clients and technical systems:

> The client will say—and it tends to be the Fortune 500 client—will say, "[w]e have a contract to do all of our e-discovery litigation with XYZ vendor." It means that they get a better price on it. XYZ knows their data systems better, so they [the client] don't need to go out and pay my hourly rates [for me] to go learn how their servers are set up and all that kind of stuff.[70]

It also appears that larger firms are using vendor platforms to further reduce costs and uncertainties of litigation through longer-term arrangements, standardization across litigation matters, and use of broader information-governance services that integrate litigation support. Echoing what other respondents told us, Chris Graham, who works at an e-discovery vendor, explained that his company has evolved to provide a wider array of information-governance services beyond e-discovery, including "development of data policies, so everything from mobile devices, social media, definitely records-retention and disposition schedules. We work on implementing those. We consult on privacy . . . And we also do e-discovery playbooks—so making sure they are ready in the event they have discovery."[71]

Indeed, e-discovery vendors are providing far more than a technical system, and this carries implications for lawyer-client relationships. Some of the larger vendors of these technical platforms are actually offering a mixed system of technical tools and humans, as Graham went on to explain:

> We have an array of products, and the client will tell us what review platform they want it to go up in. Then they will tell us if they want to do the whole review themselves. If they want just staffing, they just want some attorneys, we have a staffing arm . . . so we can give them just bodies to do review. If they want us to actually run their review for them, then we have a managed review set that will set up the workflows, do all the batching of documents, do the quality controls, give reports back, so they can set that up for them. Then

69. Interview with Samir Anand, attorney (Jan. 4, 2019).
70. Interview with Angus Martin, attorney (Jan. 02, 2019).
71. Interview with Chris Graham, e-discovery vendor consultant (Jan. 18, 2019).

> we have production environments as well where we can help them produce the documents.[72]

Consulting staff at vendors often include a range of other experts, including statisticians, linguists, and data scientists, who play an important role in how predictive coding tools are used and interpreted in the discovery process. For example, one vendor representative explained their business and staffing to us:

> [W]e do not sell AI tools. We sell AI as a service. When corporate clients come to us, they will either provide the document analysis or key document identification, having it performed by attorneys who use AI tools, or they will purchase the service that [we] provide, which really is a combination of advanced technologies. The main difference is the technologies are applied by computational linguists and computer scientists who operate these technologies in a somewhat different way than lawyers would.[73]

B. IMPROVED PERFORMANCE AND HUMAN REVIEW OF TECHNICAL SYSTEMS

While cost savings and the steep increase in the volume of material in discovery proceedings appear to be the key drivers of TAR and predictive coding tools specifically, we did encounter the standard refrain of Big Data and machine learning advocates that algorithmic systems are better—less biased, more consistent and predictable—than fallible, sometimes malicious, humans. And this sentiment came not only from technology company representatives but also from within law firms. Joe Goodman, who is a law firm litigation support manager and works closely with attorneys at his firm, reflected this sentiment:

> Yes, unequivocally, [predictive coding is] generally considered to be more accurate [than human review] because it's an algorithm. It's not a human who blinked at the wrong time or got distracted by their dog or a search term was wrong and pulled back the wrong data, those kinds of things. There's so many reasons why a human review is flawed compared to using the technology.[74]

Even when asked about training data and other factors that could influence model performance, Goodman did not waiver in his assessment of the relative accuracy of predictive coding tools compared to humans: "Those factors don't really play into it. It's a matter of comparing like populations, or

72. *Id.*

73. Interview with Ken Summers, vendor executive (Feb. 04, 2019).

74. Interview with Joe Goodman, litigation support manager at law firm (Jan. 11, 2019).

two identical populations, for human review versus algorithmic review. You're going to see greater accuracy from the algorithmic review almost every time than you would from humans."[75]

What did respondents have to say about human review of system performance? Lawyers did report using human review as a check on system outputs. However, its use was selective in ways that, if typical of practice, risked introducing a systematic bias of under-disclosure.[76] For example, Goodman, the litigation support manager, prefaced his discussion about lawyers' interactions with TAR systems with "[i]t's funny, it's almost always driven by volume."[77] He went on to explain how during the early stage of discovery, if his team is developing search terms to set an initial training set of documents, a typical conversation with attorneys would go as follows:

> [Attorneys will say] 'Here, run this group of search terms,' and . . . want to know how many documents it brings back. Then they say, 'Oh, that's too many. We've got to change the terms.' They've set the terms based on the number of documents that they've returned. Then they get the other side to agree to the search terms we're using and vice-versa.[78]

Goodman assumed this approach by the lawyers was based on "[c]ost, effort, and time."[79] He understood that this was no proper way to determine responsiveness or address the discovery principles of proportionality and defensibility: "That's usually how that goes. It's very funny, and I've never really understood this. How is it that we're determining what to review based on how many documents come back on a given search term set? Either the search terms are perpetually responsive or they're not."[80]

For their part, lawyers indicated being particularly averse to certain kinds of failures, namely the inadvertent production of privileged material. This leads post-predictive coding human reviews by attorneys to focus on documents that the system identified for production (i.e., documents scored by the predictive coding system at a probability that meets or exceeds the system's decision threshold to be classified as responsive, notwithstanding

75. *Id.*

76. Compared to lawyers, vendors reported more reliance on, and evinced a much deeper understanding of, traditional model-evaluation metrics like recall and precision. With respect to human review, they did not have objections to it, but on this, they tended to recommend the minimal amount of human review that, in their analysis, would best balance satisfying defensibility standards from the courts and saving costs on human review for their clients.

77. *Supra* note 74, at 25.

78. *Id.*

79. *Id.*

80. *Id.*

any other privileges or exceptions that might prevent disclosure). It also leads the human reviewers to give comparatively less attention, if any, to those documents *not* classified as potentially responsive (i.e., predicted negatives) or to conduct a systematic review of both groups of scored documents (i.e., both the predicted positives and predicted negatives). Respondents reported very little questioning or real review of predictive coding model performance with respect to false negatives (i.e., documents that are actually responsive but not classified as such by the predictive system).

V. IMPLICATIONS: ETHICS AND VALUES

In this Section, we discuss important implications raised by the interview data concerning professional ethics and values, the exercise of professional judgment, and the practice of law. Here, with respect to predictive coding technologies and lawyers, we focus specifically on the duty of technological competence, accountability, and lawyers' obligations of supervision when working with others involved in a case, and professionalism, disclosure, and interactions with opposing counsel.

A. LAWYER'S DUTY OF COMPETENT REPRESENTATION

As discussed above, attorneys have a professional ethical obligation to provide competent legal representation to their clients. For attorneys in most states today, that duty of competence entails keeping abreast of changes in the law and its practice, "including the benefits and risks associated with relevant technology."[81] What is happening in practice? And what do lawyers and legal technology professionals have to say about technology competence and automated decision support systems, particularly when it comes to TAR and predictive coding systems?

First, and considering the issue of technical expertise before getting to the more specific issue of lawyers' ethical duty of technology competence, our third-party TAR vendor respondents felt strongly that, compared to lawyers, they have the most technical expertise regarding information retrieval and predictive coding systems. As Ken Summers, an executive-level manager, explained to us:

> [T]his is a distinct professional domain, information retrieval. . . . It's truly a distinct professional field. I don't believe that at scale any company or any law firm or a company like ours can have truly

81. 36 states have now formally adopted the American Bar Association's 2012 revised Comment 8 to Rule 1.1 of the Model Rules of Professional Conduct. *See* Ambrogi, *supra* note 53. Other states, such as California, can impose the same or similar duties through state bar ethics opinions. *See* Cal. St. Bar Standing Comm. on Prof'l Responsibility & Conduct, Formal Op. No. 2015-193, *supra* note 56.

> two completely distinct twin core competencies. [Company name], I think, is probably today the best information retrieval company in the known universe when it comes to data analytics and litigation, investigations, etc., but we will never be a great law firm, even if we tried. It's just two distinct professional domains.[82]

Later in the interview, Summers described the average lawyer as "a lay person" who is ill-equipped to leverage "the scientific domain of search and review and information retrieval," which leads to "inefficiencies." Similarly, Paul Young, manager at an e-discovery vendor, spoke to the issue of lawyers' lack of understanding of technical features of systems and statistical concepts underlying system outputs:

> Most of them [lawyers] don't really even to go as far as to want to talk about the underlying technology[.] . . . I think statistics in general are concepts that, really, attorneys do not like. They're not familiar with or comfortable with them at all. I think anytime you're talking about defensibility and proportionality, those are generally considerations that the lawyers are familiar with. Once you start throwing things like statistics in there, saying, "In order for you to have defensible results, or in order for you to make a proportionality argument, blah, blah, blah, statistics." I think that's where a lot of the attorneys out there really shut down or they have a hard time really buying into it.[83]

How, then, are lawyers thinking about the connection between technical competence and liability? Matt Rogers, a senior attorney, articulated a general concern that lawyers may have about responsibility: "Where are the responsibilities if the platform gets screwed up? Or you make mistakes? Or you make a representation that's belied by the data? That kind of thing."[84] He went on to observe that the new arrangements between attorneys, clients, and technical expert vendors produced "decision-making friction . . . between what a [firm] wants to do, and what a client wants to do, and what the third-party provider wants to do."[85] However, going against the stereotype of risk-averse lawyers and our expectation that our lawyer respondents would point to concerns about liability risk due to inadequate understanding of black-box AI-based tools, our interviewees indicated an overall *lack* of concern about potential professional malpractice liability risk when discussing the factors driving adoption and use of these systems. Instead, expressing a general sentiment expressed by our lawyer respondents

82. Interview with Ken Summers, vendor executive (Feb. 4, 2019).
83. Interview with Paul Young, manager at e-discovery vendor (Jan. 16, 2019).
84. Interview with Matt Rogers, attorney (Jan. 7, 2019).
85. *Id.*

about their work, an attorney at a large firm who oversees the procurement of technical systems for the firm's lawyers explained:

> The attorneys at [firm] are so diligent and so focused on providing value to their clients that—well, the best legal services for the client, even if it's not value in terms of dollars and cents—no one's been worried that this is going to be a shortcut that leads to some sort of malpractice problem.[86]

B. RESPONSIBILITY AND THE DUTY TO SUPERVISE OTHERS

Even if a lawyer lacks an adequate understanding of the algorithms and models underlying TAR and predictive coding, as we discussed earlier, he or she can satisfy the ethical obligation of competent representation by associating with and supervising a sufficiently competent lawyer (within or outside the firm) and even a non-lawyer technical expert.[87] How are lawyers and third-party vendors thinking about these issues of technology competence and the duty to supervise?

Chris Graham, who is a licensed attorney but works for a TAR vendor as an e-discovery consultant, had a particularly illuminating response when asked about the extent to which he "owns the discovery process" (his phrase) in his role as consultant:

> Even though I and others on my team have decades of experience among us and we're all licensed [to practice law] in multiple states, we cannot practice law as we sit here as serving consultants. We can consult. We have to be supervised by an attorney. It means I can't just have a paralegal hire me, and there's no one else that my stuff is going through for the discovery side of things.
>
> [But] . . . we can, basically, own as much as the process as our client wants us to and control it as if we were the attorneys. We have all run these types of cases when we were practicing attorneys. At the end of the day, I need to be disclosing everything that I'm doing to an attorney so that they can satisfy their duty to supervise me, and as a barred attorney myself, it creates this weird duty to be supervised. Unlike any expert where it's not [pause]—the attorney has to understand everything I've done. They have to make sure that I'm not clearly just being reckless and doing things I shouldn't do, and if there's a big decision to be made, consulting with my

86. Interview with Chad Mankins, attorney (Aug. 22, 2018).

87. *See, e.g.*, *supra* note 56 (discussing California's ethics rules governing associating with co-counsel or technical experts for purposes of satisfying the duty of competent representation).

> client, making sure they're educated around their different options, and making a recommendation to them.[88]

Graham's response reveals some of the tensions that, in practice, the duty to supervise imposes on the attorney-client-vendor relationship. As a licensed attorney, he exhibits a keen awareness of the ethical requirements and the limitations they impose on vendor/consultant interactions with the supervising attorney (e.g., be hired by the attorney, keep the attorney informed). Yet, Graham also characterizes himself as a deep expert (e.g., "decades of experience") and able to control the e-discovery process "as if we were attorneys."

Reflecting on this shift to third-party vendors, and vendors' crucial (but often downplayed) human workers who manage discovery on a day-to-day basis, most lawyer respondents were clear that the lawyer remains responsible for errors or mistakes in discovery, even if they were made by the vendor. Steve Watson, a managing project attorney at a law firm, exemplified this sentiment: "[Y]ou cannot outsource that [responsibility] to the vendor. I know the vendors can really be helpful with the consulting work, but the lapels that the client and the court is going to grab are the firm's."[89] Similarly, Samir Anand, partner at a law firm, said: "[M]y guess is all lawyers know that they are the ones responsible. I mean, none of us go to court and say, 'Well, we used e-discovery software, so that was the problem.' "[90]

The conviction with which our lawyers proclaimed that the managing attorney is ultimately responsible for discovery does not take away from the fact that, in practice, things can get murkier. Clay Simpson, manager of legal technology and analytics at a law firm, voiced concern about law firms (not his own) where "responsibility is being outsourced to the vendor":

> Q: What's your sense of the field in general, the legal field, about knowledge of those duties, commitments, or actual practices to have those clearly defined accountability chains?
>
> A: Yeah, I'd say it's hit or miss. I'd say it's limited. A lot of folks will—a lot of other law firms will have maybe an e-discovery practice, but they're not looped in on these issues. I think it varies widely. I made the outsource-to-the-vendor point because that's what I hear a lot of my peers complain about. A lot of that responsibility is being outsourced to the vendor. It's all great if everything works perfectly. Oftentimes, a good vendor consultant can testify, even if you're being challenged, but I think if something goes really poorly, I think that's probably a bad strategy.

88. Interview with Chris Graham, e-discovery vendor representative (Jan. 18, 2019).
89. Interview with Steve Watson, attorney (Jan. 10, 2019).
90. Interview with Samir Anand, attorney (Jan. 4, 2019).

> Q: What do the vendors have to say about the issue? Is that something that they're worrying about as a liability issue?
>
> A: They're worried about it; there's no doubt about it. . . . I've talked to vendors just recently at conferences that have actually dealt with this issue. I had somebody come up to me after [a conference] who said, "I ran into the same thing where it got outsourced to me." The person there was trying to combat it and push it back to the firm, then things went poorly, and then it was just a "ball dropped" type of scenario. That's rough if the law firm itself isn't owning it, owning the project management and the AI side of it.[91]

Vendors even invoke specific strategies to protect themselves from blame and keep the lawyers responsible for legal determinations. For example, Paul Young, e-discovery vendor manager, told us that his company "tr[ies] to avoid definitively saying anything like, '[y]ou can stop now,' or '[t]his is definitely good enough.' " He explained that "that's a legal determination, and we're not the outside counsel."[92]

Finally, although not a focus of our interviews, there is a question about the extent to which attorneys may confuse contractual obligations with professional obligations. For example, when asked whether lawyers have an ethical duty to inform clients about the use of a given technology product, one law-firm respondent suggested that the contractual arrangement between corporate client and vendor addressed this concern:

> [W]henever we [law firm] contract with vendors, typically we get the client to sign the letter of engagement with the vendor directly so we don't act [as] the middle man for payment. We want the client to be on the hook to pay the vendor directly so we're out of the loop on that. They know what they're getting into. They know what they're signing up for. They know what tools are going to be used and they'll know how much it's going to cost, and they're in agreement with those terms.[93]

This assumption that service procurement will address concerns about whether or not technical choices should be discussed specifically with the client points to a risk of confusion about who is accountable for what in these triangulated relationships.

---

91. Interview with Clay Simpson, manager of legal technology at a law firm (Jan. 10, 2019).

92. Interview with Paul Young, e-discovery vendor manager (Jan. 16, 2019).

93. Interview with Joe Goodman, litigation support manager at law firm (Jan. 11, 2019).

C. INTERACTIONS WITH OPPOSING COUNSEL—TRUST, TRANSPARENCY, FAIRNESS

Finally, we address a different aspect of ethics that bears on attorney competence and fair dealing: how attorneys interact with opposing counsel during discovery. All attorney respondents expressed a preference for working with the other side to agree on the use of predictive coding. This is in line with the goal of the FRCP to encourage cooperation among parties in e-discovery.[94] Matt Rogers, an attorney who heads the e-discovery practice at his firm, reflected this preference:

> If they're [opposing counsel] looking at—you throw them your non-responses, and they say, "Hey, this is just not—you're not picking up a certain issue." [We will say,] "sorry about that, we'll pick that up." [Or they say,] "your precision is, at this level, we would like it to be higher." Maybe you agree beforehand on what it is. If the parties are being cooperative, it can be very productive, actually, to get people—I mean, you're holding down costs on both sides.[95]

However, our attorney respondents were not particularly worried about learning everything they could learn about their adversary's predictive coding system (e.g., seed-set disclosure, scoring/ranking methods, evaluation metrics). Instead, they revealed, they tend to rely on their own expertise, follow guidance from their own e-discovery vendors, and trust in their opposing counsel not to act nefariously.[96] As an e-discovery vendor manager told us, never in his career had he been asked "to explain why certain data subsets were not produced by virtue of some cutoff that left them out of the production universe."[97] This could be due to his company being "proactive" and developing comprehensive defensibility plans for their clients, as he suggested, but it could also point to insufficient technical understanding by opposing counsel.

Finally, speaking to issues of competence and reflecting the preference for agreements between parties to reduce any potential ethical predicaments, Robert Baker, attorney at a large defense firm, stated: "I think that having a

94. See discussion of Federal Rules of Civil Procedure, *supra* note 43.

95. Interview with Matt Rogers, attorney (Jan. 7, 2019).

96. Jurisdictions are split on whether, and under what circumstances, parties are required to disclose seed sets used for model training. *See* Shannon H. Kitzer, *Garbage in, Garbage out: Is Seed Set Disclosure a Necessary Check on Technology-Assisted Review and Should Courts Require Disclosure Notes*, 2018 U. ILL. J.L. TECH. & POL'Y 197 (2018). Proponents of continuous active learning (CAL), or TAR 2.0 as described earlier in the paper, may point to the seed-set disclosure issue as a reason to use TAR 2.0, as it does not require an initial set of documents for training.

97. Interview with Paul Young, e-discovery vendor manager (Jan. 16, 2019).

stipulation from the other side is pretty close to a proxy for confidence. I think [if] both sides agree to something, it's difficult for both sides to be incompetent at once, I think."[98]

Similarly, Frank Goldman, an attorney at a large plaintiff's law firm whom we expected would be more distrustful of the other side than his corporate defense counterparts, instead reinforced our finding on this issue:

> Q: I guess what I'm getting at is do you think it would be useful for there to have more clear guidance or standards about understanding the other side's TAR process and TAR system?
>
> A: More transparency I think, as a general rule, is better. I think that—I'm pausing because it's a heavy question. What should be happening is your document requests should be honestly and properly and carefully followed and answered. On some level, I care how the materials are gathered and I care what the search protocols are and I want to be really, really strategic and smart, and even if not paranoid, then very, very deliberate and careful. But if I were adjusting a dial, it wouldn't necessarily be so that I could peer into the TAR process of my adversary. It would be so that I could effectively trust and verify their discovery production.[99]

VI. ALIGNING TOOLS WITH PROFESSIONAL LOGICS

Our findings reveal a need for closer alignment of automated legal decision-making technologies, such as the predictive coding e-discovery systems that we have described here, with the professional logics of lawyers and the legal profession.[100] Such an alignment would foster not only wider adoption and deeper user understanding of these systems but would also increase public trust and the accountability of a legal profession that will continue to use these automated decision-making tools. These goals can be accomplished not only via more detailed and clearly articulated professional norms and rules—such as the duty of technological competence discussed above—but also via clearer standards, shared evaluation practices, and technical design considerations aimed at connecting these technological systems to the professional domain in which they are deployed. We briefly highlight two suggestions below as a starting point.

98. Interview with Robert Baker, attorney (Dec. 21, 2018).

99. Interview with Frank Goldman, attorney (March 8, 2019).

100. *See* Frank A. Pascale, *Professional Judgment in an Era of Artificial Intelligence and Machine Learning*, 46 BOUNDARY 2, 73–101 (2017) (contrasting the reductionist epistemology and functionalist assumptions underlying substitutive automation with the holistic epistemology of professional judgment and the conflictual, political, and contestable nature of professional work, particularly in the education and healthcare professional domains).

A. NEED FOR VALIDATION AND TESTING

First, we suggest the legal profession work to develop articulable, accessible, and consistent methods and standards for validation and testing of predictive coding tools. Our interviews indicate that not only do most lawyers lack an adequate understanding of testing and validation terms and metrics, such as recall and precision, there also does not seem to be a consistent effort to create testing schemes and datasets on which to evaluate the systems offered by TAR vendors on the market today. Instead, most vendors offer in-house metrics and validation claims as part of their marketing efforts, which of course will not be subject to the same level of benchmarking and scrutiny compared to industry-wide sets and standards.

Further, although our respondents may mention informal guidelines or validation rules of thumb, such as best practices developed by professional groups (e.g., Sedona Conference Working Groups on e-discovery issues (Working Groups 1–2, 6–7), the EDRM at Duke Law School) or in-house discovery protocols developed by vendors/consultants or attorneys themselves, they invariably tell us that they look to the more formal rules of civil procedure regarding discovery and rules of ethical and professional conduct discussed here as guideposts for the more informal governance mechanisms they employ in practice.[101] Of course, such "formal" governance mechanisms favor more general principles like "proportionality" and "defensibility" or post-hoc, interpretative evaluations provided by courts instead of articulating more technical and specific testing procedures, validation protocols, and data.

At one time, there was a Legal Track at the Text Retrieval Conference (TREC), which is sponsored by the National Institute of Standards and Technology (NIST).[102] Its stated aim was "to assess the ability of information retrieval techniques to meet the needs of the legal profession for tools and methods capable of helping with the retrieval of electronic business records, principally for use as evidence in civil litigation."[103] The conference provided a venue for shared development of datasets, identification of learning tasks—including tasks modeling discovery production in civil litigation—and evaluations of precision and recall.[104] However, the last time that the Legal Track convened at TREC was 2011, and its website indicates that it is no

101. *See* Sedona Conference, *supra* note 52; EDRM, *supra* note 50.

102. TREC LEGAL TRACK, https://trec-legal.umiacs.umd.edu/ [https://perma.cc/HN6Y-HUZV] (last visited Sept. 20, 2019) (showing little to no activity since 2011–12).

103. *See* Grossman & Cormack, *supra* note 26 (using TREC 2009 Legal Track data to conduct their study of predictive coding compared to human review in their seminal article on predictive coding).

104. *See* TREC LEGAL TRACK, *supra* note 102 (including data sets, learning tasks, evaluations, and evaluation metrics for the archive of Legal Track).

longer active. Furthermore, while the conference supported improvements in legal information retrieval techniques generally, it explicitly was not a venue for commercial product tests.[105] The closest that the legal profession comes today is with organizations like the Sedona Conference, which works to address best practices and guidelines for lawyers dealing with e-discovery issues but does not provide, to our knowledge, benchmarking tools or rigorous empirical evaluations of systems on the market.[106]

In the medical field, clinical decision support systems used to support medical judgment are subject to two forms of review: 1) explicit regulatory oversight of medical devices and 2) review by doctors and medical institutions, informed by the profession's understanding of legal-ethical duties.[107] The division of responsibility for approving which medical devices are fit for the marketplace, made by the FDA, and which tools a given medical provider chooses to use reflects an understanding that medical professionals do not have the expertise required to evaluate the performance of complex technical systems alone. At this moment, regulators[108] and doctors[109] are considering how to regulate clinical decision support systems—

105. *See* TREC STATEMENT ON PRODUCT TESTING AND ADVERTISING (Apr. 9, 2019), https://trec.nist.gov/trec.disclaim.html [https://perma.cc/97XL-HANF].

106. *See* Sedona Conference, *supra* note 52.

107. For our purposes, automated clinical decision support systems relying on machine learning to aid medical doctors in making decisions are a useful comparison for the predictive coding systems in law that we study below.

108. 21st Century Cures Act, Pub. L. No. 114-255 (2016). In December 2016, President Obama signed into law the 21st Century Cures Act. Section 3060(a) of the Cures Act added a new subsection to the Food, Drug, and Cosmetic Act (FDCA) that excludes from the Food and Drug Administration's (FDA) medical-device regulations and approval processes "software function" that meets the following conditions: 1) not intended to acquire, process, or analyze a medical image or a signal from an in vitro diagnostic device or a pattern or signal from a signal acquisition system; 2) intended for the purpose of displaying, analyzing, or printing medical information about a patient or other medical information (such as peer-reviewed clinical studies and clinical practice guidelines); 3) intended for the purpose of supporting or providing recommendations to a health care professional about prevention, diagnosis, or treatment of a disease or condition; and, 4) intended for the purpose of enabling such health care professional to independently review the basis for such recommendations that such software presents so that it is not the intent that such health care professional rely primarily on any of such recommendations to make a clinical diagnosis or treatment decision regarding an individual patient. 21 U.S.C. § 360(j)(o)(1)(E)(i)–(iii) (2016).

109. *See, e.g.*, Emily L. Evans & Danielle Whicher, *What Should Oversight of Clinical Decision Support Systems Look Like?*, 20 AM. MED. J. ETHICS 857 (2018) (arguing that while using a clinical decision support system may not be a research activity under the Common Rule, its use requires more ethical and regulatory oversight than clinical practice and proposing a framework that sets out conditions governing use, ongoing monitoring of data quality, processes for developing and validating algorithms, and protections for patient data); Nicole Martinez-Martin et al., *Is It Ethical to Use Prognostic Estimates from Machine Learning to Treat Psychosis?*, 20 AM. MED. J. ETHICS 804 (2018) (providing an example of how the profession is

particularly those that rely on machine learning. These two regulatory forces provide different types of expertise that can collectively work to align machine-learning tools with the fields' decision-making processes. Regardless of how exactly clinical decision support systems are governed by regulatory bodies like the FDA, other factors—e.g., professional licensing requirements, ethical duties, tort-based malpractice liability principles, and doctors' own conceptions of themselves as users of these technologies—will shape clinical decision support tools and the conditions of their adoption and use. The exact contours of these various ethical and legal obligations are still emerging, but professionals and professional associations are keenly aware of the need to actively shape these tools to serve the needs of the medical field.[110] They are pushing for tools that are interpretable by medical professionals and used under conditions that support "epistemically responsible" knowledge production and behavior.[111]

Our research of the legal sector reveals a profession struggling to evaluate increasingly complex tools without requisite expertise and systematic and shared methods. The gatekeepers and gatekeeping tools historically relied upon are insufficient to oversee the influx of predictive machine learning systems into legal practice. Reliance on professional rules and court approval is untenable.[112] The legal profession needs to develop new governance models that enlist appropriate technical experts in evaluating systems that support professional cognitive work. Setting aside the question of whether the creation of a separate body charged with evaluating TAR and other tools that support, augment, or replace professional decision making is necessary, there is at least a pressing need for shared methods and standards for validation and testing of predictive coding tools.

B. TOWARD CONTESTABILITY AS A FEATURE OF DECISION-MAKING TOOLS

Assuming for the moment that the profession heeds our call and develops processes for evaluating predictive coding tools by those with

---

grappling with such machine learning-based decision support systems); *Augmented intelligence in health care H-480.940*, AM. MED. ASS'N https://policysearch.ama-assn.org/policyfinder/detail/augmented%20intelligence?uri=%2FAMADoc%2FHOD.xml-H-480.940.xml [https://perma.cc/FW22-VU9P] (last visited Sept. 20, 2019) (setting out principles to guide development and use of AI).

110. Danton S. Char et al., *Implementing Machine Learning in Health Care — Addressing Ethical Challenges*, 378(11) NEW ENGLAND J. MED. 981, 981 (2018).

111. Simon, *supra* note 9, at 145.

112. Dana Remus & Frank Levy, *Can robots be lawyers: Computers, lawyers, and the practice of law*, 30 GEO. J. LEGAL ETHICS 501, 556 (2017) (arguing for the adoption of "more effective regulatory structures that draw upon both legal and technical expertise, while protecting both clients and the values of our legal system").

requisite expertise, those systems will still require run-time configuration. Aligning a TAR system with discovery needs in a particular case, or with respect to a particular set of documents, requires exposing relevant aspects of system design and, where possible, opening them up to exploration and configuration. Professionals appropriate technology differently, employing it in everyday work practice, as informed by routines, habits, norms, values, and ideas and obligations of professional identity. Appropriate handoffs to, and collaborations with, decision-support systems demand that they reflect professional logics and provide users with the ability to understand, contest, and oversee decision making. Technical design should seek to put professionals and decision support systems in conversation, not position professionals as passive recipients of system wisdom who must rely on out-of-band mechanisms to challenge them. For these reasons, calls for explainability[113] fall short and should be replaced by governance approaches that promote contestable systems. This requires attention to both the information demands of professionals—inputs, decisional rules, etc.—and processes of interaction that elicit human expertise and allow humans to elicit information about machine decision making.

To foster user engagement and understanding, and to surface the values implicated by TAR systems and decisions, we embrace the design principle of "contestability."[114] As we have described the concept elsewhere,

113. Compared to "explainability" as a value goal for system design, contestability is a more active and dynamic principle. Where the passivity of "explainable" algorithmic systems imagines engagement, reflection, and questioning as out-of-band activities—via exception handling, appeals processes, etc.—contestable systems are designed to foster active, critical engagement within the system. Explanations, as reflected in policy debates and the majority of research on interpretable systems, are also typically viewed as static—focused on conveying a single message. Ashraf Abdul et al., *Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda*, *in* PROC. INT'L CONF. ON HUM. FACTORS IN COMPUTING SYSTEMS 1 (2018) (reviewing the explainable AI literature and observing that researchers in this community tend to produce static explanations).

114. *See* Tad Hirsch et al., *Designing Contestability: Interaction Design, Machine Learning, and Mental Health*, *in* DES INTERACT SYST CONF. 95, 98 (2017) (setting out contestability as a design objective to address the myriad ethical risks posed by the potential reworking of relationships and redistribution of power caused by the introduction of machine-learning systems. In their example, they explain how a machine learning-based assessment and training tool for psychotherapists could be used as a "blunt assessment tool" of management). They offer three lower-level design principles to support contestability: 1) improving accuracy through phased and iterative deployment with expert users in environments that encourage feedback; 2) heightening legibility through mechanisms that "unpack aggregate measures" and "trac[e] system predictions all the way down" so that "users can follow, and if necessary, contest the reasoning behind each prediction," 3) identifying "aggregate effects" that may imperil vulnerable users through mechanisms that allow "users to ask questions and record disagreements with system behavior" and engage the system in self-monitoring. *Id.* Together, these design principles can drive active, critical,

contestability refers to mechanisms for users to understand, construct, shape, and challenge model predictions.[115] It is a particularly important system quality where the goal is for predictive algorithms to enhance and support human reasoning, such as decision-support systems and systems that aid users in evaluative cognitive tasks. A wide array of empirical studies provide evidence that interactive, contestable systems advance individual user understanding.[116] In addition, such systems can not only improve user understanding and use of a system but also enable users to provide deep and useful feedback to improve algorithms.[117]

Contestable design would allow lawyers to more dynamically explore and interact with TAR systems. In particular, the responsive and dynamic tailoring of continuous active learning-based predictive coding systems, combined with rich feedback and interaction with professional experts, could produce decisions that support "epistemically responsible" knowledge production.[118] Contestability spreads the production of knowledge across humans and machines. Indeed, systems designed for contestability invite engagement rather than delegation of responsibility, which aligns well with regulatory and liability principles that seek to keep humans in the loop. They can foster engagement through both the provision of information about system inputs, reasoning, and outputs, and through an interactive design that encourages exploration and querying. In other words, contestability makes algorithmic systems knowable to lawyers, responding to their need (and ethical duty of technological competence) to understand the tools one uses while simultaneously responding to the societal need to ensure that tools are fit for purpose. Contestable design thus contributes to the creation of governance models that support epistemically responsible behavior[119] and encourages shared reasoning about the appropriateness of algorithmic systems' behavior.

---

real-time engagement with the reasoning of machine-learning system inputs, outputs, and models.

115. Deirdre K. Mulligan et al., *Contestability: From Explanations to Engagement with AI*, *in* AFTER THE DIGITAL TORNADO: NETWORKS, ALGORITHMS, HUMANITY (Kevin Werbach ed., 2019).

116. Saleema Amershi et al., *Power to the People: The Role of Humans in Interactive Machine Learning*, 35 AI MAG. 105 (2014).

117. Simone Stumpf et al., *Toward Harnessing User Feedback for Machine Learning*, *in* 12TH INT'L CONF. ON INTELLIGENT USER INTERFACE 82 (2007); *see also* S. Stumpf et al., *Interacting Meaningfully with Machine Learning Systems: Three Experiments*, 67 HYPERCONNECTED INT'L J. HUM. COMPUTER STUD. 639 (2009).

118. Simon, *supra* note 9, at 145.

119. *Id.*

VII. CONCLUSION

The introduction and increasing popularity of predictive coding systems is reshaping the legal profession in significant ways. Far from being just a new tool in "normal practice," we found that predictive coding—and, perhaps, the broader set of complex, ML-based legal technologies entering the profession—has brought new entities and technical experts into the legal services ecosystem who are mediating the relationship between lawyers and clients. This raises old questions, such as those about contract attorneys and outsourcing of legal work, but in slightly new forms and involving new parties. The result is a reconfiguration of social relations and new power dynamics, specifically 1) new kinds of professionals who have the training and expertise to build and use the tools in ways few lawyers do, and 2) a new tool for cost containment by corporate clients.

Our research reveals that more work needs to be done to address potential blind spots at the intersection of professional governance (via rules of professional ethical conduct) and legal decision-support technologies. Lawyers are reliant not only on "black box" technical tools but also on other experts. Just in our case of predictive coding for e-discovery, lawyers relied on non-lawyer support staff and vendor judgment for a variety of tasks: system selection (reliant on vendors and, for some, on in-house litigation support staff for early testing), configuration (reliant on in-house technical experts or vendors), and model-testing and evaluation (we found no real standards or benchmarking). This points to the need to rethink how tools are tested and evaluated before they are unleashed into the field. It also points to the need for more education and training to ensure lawyers better understand ML-based decision support systems. Finally, even with better evaluation and testing of tools, in order to appropriately support lawyers on a given discovery task, such tools must be interpretable and configurable—that is, contestable. Lawyers, and the legal profession more generally, should take heed of the growing reliance on machine learning systems for professional work and develop clearer rules, standards, design features, and procedures governing the procurement, deployment, and, crucially, user understanding, of automated legal decision-making technologies.

IS TRICKING A ROBOT HACKING?

Ivan Evtimov,[†] *David O'Hair,*[††] *Earlence Fernandes,*[†††] *Ryan Calo*[‡] *& Tadayoshi Kohno*[‡‡]

DOI: https://doi.org/10.15779/Z38M32N99Q

† Ph.D Student, Paul G. Allen School of Computer Science & Engineering, University of Washington.

†† Intellectual Property Attorney, Knobbe Martens.

††† Assistant Professor, Department of Computer Sciences, University of Wisconsin-Madison.

‡ Lane Powell and D. Wayne Gittinger Associate Professor, School of Law, University of Washington.

‡‡ Professor, Paul G. Allen School of Computer Science & Engineering, University of Washington. The authors of this essay represent an interdisciplinary team of experts in machine learning, computer security, and law. The authors would like to thank the following people for their insightful contributions, comments, and suggestions: Ashkan Soltani, Michael Froomkin, Jesse Woo, Amanda Levendowski, Jason Schultz, Aaron Snoswell, Ram Shankar Siva Kumar, Natalie Chyi, Amit Elazari, Jennifer Rogers, and Dallas Card. We thank Jesse Woo for furnishing the example of the Budapest Convention on Cybercrime. Thank you to the Hewlett Foundation, MacArthur Foundation, and Microsoft for its support of the Tech Policy Lab and this work.

TABLE OF CONTENTS

I. INTRODUCTION

The term "hacking" has come to signify breaking into a computer system.[1] Lawmakers crafted penalties for hacking as early as 1986, supposedly in response to the movie *War Games* from three years earlier in which a teenage hacker gained access to a military computer and nearly precipitated a nuclear war.[2] Today, a number of local, national, and international laws seek to hold hackers accountable for breaking into computer systems to steal information or disrupting their operations. Other laws and standards have incentivized private firms to use best practices in securing computers against attacks.

1. We acknowledge that there is a second, classic, definition of "hacking," which refers to deep technical explorations of computer systems without malice. *See* INTERNET USERS' GLOSSARY (Jan. 1993), https://tools.ietf.org/html/rfc1392 [https://perma.cc/ZG5F-VL8T]. This definition contrasts "hacking" to "cracking." However, we use the more contemporary definition of hacking here.

2. H.R. REP. NO. 98-894, at 3696 (1984).

The landscape has shifted considerably from the 1980s and the days of dial-ups and mainframes. Most people in 2019 carry around in their pockets the kind of computing power available to the United States military at the time of *War Games*. People, institutions, and even everyday objects connect with each other via the Internet. Driverless cars roam highways and city streets. Yet, in an age of smartphones and robots, the classic paradigm of hacking—in the sense of unauthorized access to a protected system—has persisted. All of this remains a challenge for legal institutions.

However, in addition to the current challenges, a new set of techniques aimed not at breaking into computers, but at manipulating the increasingly intelligent machine learning models that control them, may force the law and legal institutions to reevaluate the very nature of hacking. Three of the authors have shown, for example, that it is possible to use one's knowledge of a system to fool a machine learning classifier, such as the classifiers one might find in a driverless car, into perceiving a stop sign as a speed limit.[3] Other machine learning manipulation techniques build secret blind spots into learning systems or reconstruct the private data that goes into model training.[4]

The unfolding renaissance in artificial intelligence (AI), coupled with an almost-parallel discovery of considerable vulnerabilities, requires a reexamination of what it means to "hack," i.e., to compromise a computer system. The stakes are significant. Unless legal and societal frameworks adjust, the consequences of misalignment between law and practice will result in (1) inadequate coverage of crime, (2) missing or skewed security incentives, and (3) the prospect of chilling critical security research. This last consequence is particularly dangerous in light of the important role researchers play in revealing the biases, safety limitations, and opportunities for mischief that the mainstreaming of artificial intelligence may present.

This essay introduces the law and policy community, within and beyond academia, to the ways adversarial machine learning (ML) alters the nature of hacking, and with it, the cybersecurity landscape. Using the Computer Fraud and Abuse Act of 1986 (CFAA)—the paradigmatic federal anti-hacking law—as a case study, we hope to demonstrate the burgeoning disconnect between law and technical practice. And we hope to explain the stakes if we fail to address the uncertainty that flows from hacking that now includes tricking.

3. *See* Kevin Eykholt et al., *Robust Physical-World Attacks on Deep Learning Visual Classification*, IEEE/CVF CONF. ON COMPUTER VISION & PATTERN RECOGNITION 1625, 1626 (2018).

4. *See* Ramya Ramakrishnan et al., *Robust Physical-World Attacks on Deep Learning Visual Classification* (Cornell Univ., Working Paper No. 5, 2018), https://arxiv.org/pdf/1707.08945v5.pdf [https://perma.cc/EHL6-4CCB].

The essay proceeds as follows. Part II provides an accessible overview of ML. Part III explains the basics of adversarial ML for a law and policy audience, laying out the current set of techniques used to trick or exploit AI. This essay is the first taxonomy of adversarial ML in the legal literature (though it draws from prior work in computer science).[5]

Part IV describes the current legal anti-hacking paradigm and explores whether it envisions adversarial ML. The question is difficult and complex. Our statutory case study, the CFAA, is broadly written and has been interpreted expansively by the courts to include a wide variety of activities, including overwhelming a network with noise and even violating a website's terms of service. Yet, when we apply the CFAA framework to a series of hypothetical examples of adversarial ML grounded in research and real events, we find that the answer of whether the CFAA is implicated is unclear. The consequences of certain adversarial techniques cause them to resemble malicious hacking, and yet the techniques do not technically bypass a security protocol as the CFAA envisions.

Part V shows why this lack of clarity represents a concern. First, courts and other authorities will be hard-pressed to draw defensible lines between intuitively wrong and intuitively legitimate conduct. How do we reach acts that endanger safety—such as tricking a driverless car into mischaracterizing its environment—while tolerating reasonable anti-surveillance measures—such as makeup that foils facial recognition—when both leverage similar technical principles, but produce dissimilar secondary consequences?

Second, and relatedly, researchers testing the safety and security of newer systems do not always know whether their hacking efforts may implicate federal law.[6] Moreover, designers and distributors of AI-enabled products will not understand the full scope of their obligations with respect to security. Therefore, we join a chorus of calls for the government to clarify the conduct it seeks to restrict while continuing to advocate for an exemption for research aimed at improvement and accountability. We advance a normative claim that the failure to anticipate and address tricking is as irresponsible or "unfair" as inadequate security measures in general.

We live in a world that is not only mediated and connected, but increasingly intelligent. Yet that intelligence has limits. Today's malicious actors penetrate computers to steal, spy, or disrupt. Tomorrow's malicious actors may also trick computers into making critical mistakes or divulging the private information

5. *See infra* Part III.

6. Our focus is on the CFAA but, as we acknowledge below, other laws such as the Digital Millennium Copyright Act (DMCA) also establish penalties for unauthorized intrusion into a system. The DMCA, however, has an exception for security research.

upon which they were trained. We hope this interdisciplinary project begins the process of reimagining cybersecurity for the era of artificial intelligence and robotics.

II. MACHINE LEARNING

AI can best be understood as a set of techniques aimed at approximating some aspect of human or animal cognition.[7] It is a long-standing field of inquiry that, while originating in computer science, has since bridged many disciplines.[8] Of the various techniques that comprise AI, a 2016 report by the Obama White House singled out ML as particularly impactful.[9] Underpinning many of the most impactful instantiations of AI, ML refers to the ability of a system to improve performance by refining a model.[10] The approach typically involves spotting patterns in large bodies of data that in turn permit the system to make decisions or claims about the world.[11] The process subdivides into two stages: training and inference.[12] During training, available data is used as input to generate a model that is oriented toward a particular objective such as fraud detection.[13] Then, during inference, researchers deploy the trained model to make claims or predictions about previously unseen data, such as new bank transactions.[14]

A. DEEP LEARNING

A particularly promising ML training technique is referred to as "deep learning." Until recently, few good ML algorithms could rival human performance on common benchmarks. For instance, identifying an object in a picture or finding out to whom a facial image belongs represented a high challenge for computers in the past.[15] However, a confluence of greater

7. Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. DAVIS L. REV. 399, 404 (2017).

8. *See id.* at 404–05.

9. *See* EXEC. OFFICE OF THE PRESIDENT, ARTIFICIAL INTELLIGENCE, AUTOMATION, AND THE ECONOMY (2016), https://obamawhitehouse.archives.gov/blog/2016/12/20/artificial-intelligence-automation-and-economy [https://perma.cc/3RCA-NMAT] [hereinafter White House Artificial Intelligence Report].

10. *See* SHAI SHALEV-SHWARTZ & SHAI BEN-DAVID, UNDERSTANDING MACHINE LEARNING: FROM THEORY TO ALGORITHMS 1–7 (2014). We are not committed in any deep sense to the idea that ML falls within, rather than adjacent to, AI. However, we adopt for purposes of this essay the conventional frame that ML is a form of AI.

11. *Id.*

12. *Id.*

13. *Id.*

14. *Id.*

15. *See* White House Artificial Intelligence Report, *supra* note 9, at 6.

availability of large datasets, advances in parallel computing, and improvements in processing power have helped deep learning models achieve human-level or better performance.[16] Subsequently, researchers applied deep learning in a host of other areas, which led to the past decade's explosion in deep learning applicability and use.

Deep learning involves the distillation of information presented in a complex format (for instance, pixels) down to easily interpretable labels (for instance, an object category) by layering the processing of information.[17] For example, on the first layer of image processing, a deep-learning algorithm might attempt to detect boundaries between objects in an image. A subsequent layer might be responsible for grouping these boundaries into shapes, and deeper layers might assign higher-level meanings to the shapes. Eventually, the final layer might translate the outputs of all of those layers into a simple concept, such as the category of the object in the image.

Deep learning itself is not a new concept; indeed, many ML models before deep learning attempted to do just what deep learning does, using layers handcrafted by researchers.[18] For instance, to classify faces, scientists would specify how to process the images by trying to define which regions of the face were important for predicting identity.[19] One of deep learning's innovations was to let each computational layer adjust itself automatically based on the training data.[20] This is achieved by mathematically defining how close the output of the final computational layer is to what is desired, and how to update the intermediate layers so that the overall output gets closer to the target. Thus, with enough data and time, the model will strive to get better at outputting the label "dog" for all images of dogs.[21]

B. TRAINING DATA

The large amount of training data that has recently been made readily available represents one of the key factors that has allowed ML techniques generally, and deep learning models in particular, to become useful in a broad variety of applications. Training data can be provided to ML algorithms in

16. *Id.*
17. *See* IAN GOODFELLOW ET AL., DEEP LEARNING 1 (2016).
18. *See id.* at 2.
19. *See id.* at 3.
20. *See id.* at 5.
21. Furthermore, matrix multiplications represent the internals of deep learning models. For many decades before deep learning took off, computer scientists had been studying how to make those operations execute quickly and in parallel. Thus, deep learning also has the benefit of naturally parallelizing computations at a time when the performance of non-parallel computing power flattened out.

many ways, since many datasets are created in laboratories where conditions can be specified precisely. For example, when building a face-recognition training set, taking images in a lab can provide exact details of the position of the subject's head, camera-exposure settings, lighting conditions, etc.

However, this is not always practical, especially for data-hungry models such as today's deep learning frameworks. These algorithms need many more precisely labeled examples than individual researchers could possibly generate. Thus, in many applications, the training set and its labels are automatically generated or crowdsourced.[22] For instance, to generate an image-recognition dataset, one might simply select all images that come up in a Google Image search for the categories of interest. If a researcher wanted to build a classifier of car models, they would search for each model on Google Images and download the resulting pictures. Alternatively, one could collect a lot of images with unknown labels and then ask online volunteers to label them. A third option is to attempt to infer the labels from user activity. For example, to generate a text-prediction dataset from words typed in a keyboard, one might simply look at what a user chooses to type next.

Sources of training datasets turn out to be extremely important in channeling the social impacts of ML. Whether created synthetically in a lab, purchased from a vendor, or scraped from the Internet, the dataset a model encounters during its training phase will dictate its performance at the application or interference phase.[23] If the training data is historically sourced, it will contain the biases of history. If the training data is not representative of a particular population, the system will not perform well for that population. A generation of contemporary scholars is adding to the existing literature concerning the ways machine encode and perpetuate bias.[24] As our focus is on computer security, we refer to the conversation here only to acknowledge its importance.

C. LIMITATIONS OF MACHINE LEARNING

For all the impressive results that deep learning models have achieved, scholars and their ML models still suffer from a host of limitations that has

22. *See* GOODFELLOW ET AL., *supra* note 17, at 17 (discussing the benchmark datasets that are used to train and evaluate deep learning models).

23. *See generally* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018) (discussing the various implicit biases that shape AI and ML).

24. *See, e.g.*, *ACM Conference on Fairness, Accountability, & Transparency (AMC FAT*)*, ACM FAT Conf., https://fatconference.org/index.html [https://perma.cc/26GP-Q6G2] (last visited Oct. 10, 2019).

been well studied by statisticians and computer scientists.[25] We focus our work on the failures of these models under the threat of active adversaries, but we include a brief discussion of these well-documented shortcomings for context. We draw a distinction between those problems that are well studied in the core of ML research, and adversarial ML, which has only recently begun to take the center stage.

One limitation is that it is hard to know how "good" any given ML model is. Measuring performance itself is not a straightforward process, as many different scores exist that vary significantly depending on application.[26] In fact, what we think of as "accuracy" in common speech (i.e., the percentage of correct answers an algorithm produces to a set of questions) often poorly measures how well a model performs. Having an accurate system is not enough unless the system captured the correct target. For example, an accuracy metric is especially inappropriate in medical diagnosis systems, as it does not account for how often a disease appears in the population. Thus, a naive algorithm for a rare disease that simply classifies every case it sees as "not sick" would have fairly high accuracy score—since the disease is rare, most cases the algorithm sees will rightly be classified as "not sick." However, this would be a very bad algorithm, as it would never predict when a person actually has the disease. Thus, ML researchers have to be careful to choose metrics that capture such tradeoffs in performance and make it hard for models to "cheat."

Even when a proper evaluation metric is chosen, there are many reasons that may explain why a ML model might not do well on it: the model might not be given the right features, it might not have seen enough data, or it might not have enough capacity to capture the true complexities of the underlying process it tries to predict. Imagine someone was trying to build a model to try to predict the outcome of a presidential race based on the names of the candidates' sixth-grade math teachers. However sophisticated the model, it will not be able to properly account for the complex dynamics of national elections because it did not receive the right signal. Similarly, if a pollster asked only a single person in each state how they will vote, no model would be able to spit out a meaningful prediction because it does not have the right data. Deep learning is similar. A neural network given only images of sheep will likely to

25. *See id.*; MILES BRUNDAGE ET AL., THE MALICIOUS USE OF ARTIFICIAL INTELLIGENCE: FORECASTING, PREVENTION, AND MITIGATION (2018), https://maliciousaireport.com/ [https://perma.cc/NK5Z-HLS7].

26. *See* Tavish Srivastava, *7 Important Model Evaluation Error Metrics Everyone should know*, ANALYTICS VIDHYA (Feb. 19, 2016), https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/ [https://perma.cc/L7WW-QEHL] (discussing just one of many technical details of the wide variety of accuracy metrics and their applications scenario).

predict that all white furry animals—even if they are goats—are sheep. A shallow network might not be able to tell the difference between different breeds of sheep and different makes of cars at the same time.

Such failures might seem obvious and are likely to be detected by a poor performance on the test metric. However, a good performance even on a well-chosen metric does not inoculate a model from failures in the real world. Notably, it is easy to make ML models "overfit."[27] In such a scenario, a model captures irrelevant relationships in the underlying data and learns to make its predictions based on those, instead of uncovering the true "signal." For instance, a model that only analyzes blue flowers during training and is only tested on blue flowers might focus only on the color, and thus learn to classify only blue objects as flowers. Since the test set contains no flowers of other colors, the model would score well, but fail in the real world where non-blue flowers exist. Thus, researchers need to ensure that their training and test sets are properly balanced and include a large enough sample of relevant features. To detect overfitting, benchmark holders will keep the evaluation dataset hidden from the model developers until a final version of the model is presented.[28]

Furthermore, the power of deep learning to perform well comes at a cost. Due to the large number of parameters, deep learning models are hard to interpret. While it is trivial to look at the matrices a training set has generated, it is not clear what role these matrices are playing in the model's computation. Therefore, it is not easy to explain what any intermediate layer is doing or how it is contributing to the overall prediction. Generating greater clarity around deep learning remains an active area of computer science research.

III. ADVERSARIAL MACHINE LEARNING

As we have seen, there are many reasons why ML algorithms can fail "naturally." Statisticians have known about these possible pitfalls almost for as long as statistics has been around.[29] By contrast, a relatively new area of study indicates that deep learning systems remain vulnerable to a new class of problems when an adversary actively attempts to tamper with them or interact

27. *See* SHALEV-SHWARTZ & BEN-DAVID, *supra* note 10, at 33–42.

28. *See* KAGGLE COMPETITIONS, https://www.kaggle.com/competitions [https://perma.cc/68XV-N9RE] (last visited Oct. 10, 2019) (demonstrating how a platform can be used to host competitions for models that enforces this policy).

29. *See* SHALEV-SHWARTZ & BEN-DAVID, *supra* note 10, at 33–42.

with them maliciously.[30] As such, they present a particularly interesting challenge for the current legal framework and we center our discussion on these new techniques. Researchers to date have identified three main approaches to "adversarial" ML: (1) fooling a trained classifier or detector into mischaracterizing an input in the inference phase, (2) skewing the training phase to produce specific failures during inference, and (3) extracting the (sometimes sensitive) underlying data from a trained model.[31] We discuss each of the three approaches in turn.

A. FOOLING, POISONING, AND EXTRACTING: ADVERSARIAL ML SINCE 2013

The first approach involves "fooling" a machine. Seminal work from 2013 by Szegedy et al. discovered that in the domain of image recognition, changing only a few pixels of an image causes the model to predict a wrong label.[32] Subsequent work showed that even more sophisticated models contained vulnerabilities to such human-imperceptible "adversarial examples," and they provided powerful algorithms to find these malicious inputs.[33] Researchers also established that attackers have some latitude in picking how the model they target will misbehave.[34] An adversary could, for example, select a target class to which they will send the model's prediction of the adversarial inputs.[35] For instance, an adversary could make a warning label appear as a particular message of their choosing, such as an expiration date.

Other computer scientists discovered that adversarial examples also transfer across models performing the same task.[36] Thus, attackers could generate malicious inputs for a proxy classifier and use them to cause failure in other similar systems. For instance, Liu et al. demonstrated that one could take images of various objects, add adversarial noise to cause the image to fool a publicly accessible system, and then use the same perturbed images in a commercial image recognition service that had not been analyzed. The new,

30. By "adversary," we refer to an individual, group, or agent who intends to compromise the system, i.e., stands in an adversarial relationship with the system and its developers.

31. *See* Nicolas Papernot et al., *Towards the science of security and privacy in machine learning*, ARXIV (Nov. 11, 2016), https://arxiv.org/abs/1611.03814 [https://perma.cc/UG4B-JUCU].

32. *See* Christian Szegedy et al., *Intriguing properties of neural networks*, ARXIV (Dec. 21, 2013), https://arxiv.org/abs/1312.6199 [https://perma.cc/5UKG-Z4W5].

33. *See* Papernot et al., *supra* note 31.

34. *See id.*

35. *See id.*

36. *See id.*

unanalyzed model still predicted "veil" for an image of a dog based on the altered image.[37]

Finally, a growing body of work focuses on how to produce physical adversarial examples. For instance, two recent high-profile papers demonstrated that wearing specifically crafted glasses can trick face recognition systems[38] and that applying adversarial stickers to a road sign can cause the sign to be misinterpreted by an autonomous vehicle's image classifier.[39] Similar work also exists in the audio domain. Carlini and Wagner demonstrated that audio can be perturbed in a way not obvious to humans, but causes a different classification by a smart assistant.[40]

It must be noted that the attacks discussed so far happen *after* the model is trained and after it has been deployed, i.e., in the inference phase. The attacker can execute those attacks without interfering with the training procedure, simply by presenting the model with modified inputs.[41] However, an attacker needs to know precisely how the model she is attacking, or at least how a similar model, works.

Unlike classical security vulnerabilities such as buffer overflows, where extra data overflows into and corrupts an adjacent memory space in a computer program, these problems exist independently of who created the model. Regardless of who wrote the software for the instantiation of the targeted neural network or where they sourced their training data from, almost *any* deep learning model seems to be susceptible to adversarial examples.[42] Despite intensive research in this area since 2013, most attempts at hardening deep learning models have failed and the technical literature has expanded to include attacks on neural networks applied outside of classical computer vision tasks.[43]

Another set of attacks focuses on interfering with or "poisoning" model training. An adversary who could tamper with the training data can, in theory, compromise the model in any arbitrary way. For instance, the adversary could label all pictures of rabbits in the training set as pictures of dogs. A model will

37. Yanpei Liu et al., *Delving into Transferable Adversarial Examples and Black-box Attacks*, ARXIV (Nov. 8, 2016), https://arxiv.org/abs/1611.02770 [https://perma.cc/9PUX-QTLK].

38. MAHMOOD SHARIF ET AL., ACCESSORIZE TO A CRIME: REAL AND STEALTHY ATTACKS ON STATE-OF-THE-ART FACE RECOGNITION (2016).

39. *See* Eykholt et al., *supra* note 3.

40. Nicholas Carlini & David A. Wagner, *Audio Adversarial Examples: Targeted Attacks on Speech-to-Text*, ARXIV (Jan. 5, 2018), https://arxiv.org/abs/1801.01944 [https://perma.cc/48W9-GVYA].

41. *See id.*; SHARIF ET AL., *supra* note 38; Eykholt et al., *supra* note 3.

42. *See* Papernot et al., *supra* note 31.

43. "Hardening" refers to making the system more robust against attacks.

then naturally learn that dogs look like rabbits. Similarly, the adversary could be more subtle and train the model so that every picture of a rabbit with a particular patch of hair gets classified as a dog. However, the adversary need not control the training set or even its labels to "backdoor" an error into the model in this way.[44] One recent work demonstrated that an adversary with full access to the trained model, i.e., white-box access, can build a "trojan trigger."[45] This trigger would only cause misclassification if it is presented to the model, but will not otherwise affect the performance of the model.[46] This could become problematic for models, like an image-based search engine, distributed online or fully trained by a third party as a service.[47]

A third type of attack on deep-learning models seeks to compromise the privacy of data contained within the training set.[48] In this type of attack, an adversary needs to obtain the full model (its internal structure and weights). The attacker can then seek to either infer membership of particular individuals or reconstruct the training data. For instance, a naive text-prediction model, incorporated in a smartphone keyboard, could be inverted to extract sensitive data a user has typed in the past, such as a Social Security Number, date of birth, or private messages in an otherwise end-to-end encrypted messaging app such as Signal.[49] It is generally possible to protect against such attacks by employing a mathematical technique known as differential privacy.[50] At a high level, this technique allows one to add noise to the data in a way that preserves its useful properties for the whole dataset, but makes it hard for adversaries to reveal information about individual members.[51] However, research remains ongoing on the performance tradeoffs when employing this protective technique.[52]

---

44. *See supra* Section III.A.

45. See Yingqi Liu et al., *Trojaning Attack on Neural Networks*, Network & Distributed Syss. Security (NDSS) Symp. (Feb. 18, 2018), https://www.cs.purdue.edu/homes/ma229/papers/NDSS18.TNN.pdf [https://perma.cc/E98P-NJJ3].

46. *See id.*

47. For example, Amazon Rekognition is such a service that can train a model for certain computer vision tasks based on a dataset provided by the client. *See* AMAZON REKOGNITION, https://aws.amazon.com/rekognition/ [https://perma.cc/LDG7-R744] (last visited Oct. 15, 2019).

48. *See, e.g.*, Reza Shokri et al., *Membership Inference Attacks against Machine Learning Models*, ARXIV (Oct. 18, 2016), https://arxiv.org/abs/1610.05820 [https://perma.cc/GXZ7-GR4T].

49. *See id.* While this particular text-extraction approach is possible in theory, we previously discussed a model-extraction approach based on an image model.

50. *See* MATTHEW FREDRIKSON ET AL., PRIVACY IN PHARMACOGENETICS: AN END-TO-END CASE STUDY OF PERSONALIZED WARFARIN DOSING (2014).

51. *See* CYNTHIA DWORK ET AL., DIFFERENTIAL PRIVACY: A PRIMER FOR THE PERPLEXED 11 (2011).

52. *Id.*

B. LIMITATIONS OF ADVERSARIAL ML

Most applications of adversarial ML today are limited to academic proofs of concept and do not necessarily reflect current vulnerabilities in deployed systems. In the case of adversarial examples, deployed systems will likely employ some pre- or post-processing to their models to detect and filter adversarial examples (although no defense has worked to date).[53] In addition, no adversarial examples have been shown to defeat multiple different models at the same time. For instance, a self-driving car that perceives adversarial stop signs that an image classifier mistakes for speed limit signs might still detect the sign correctly via its light detection and ranging (LiDAR) technology.[54]

Furthermore, generally the most powerful attacks currently occur only with full "white box" knowledge of the models that are targeted. Although these models are proprietary, computer science research points out that such proprietary restriction does not always prevent attacks because adversarial examples designed for one model can often attack similar, unknown models as well.[55] But attacks that do transfer across models generally include much higher distortions that might be noticeable to humans.[56] Similar limitations exist for model inversion attacks.[57]

While the space of attacking ML models is still technologically young, later we will present several case studies that might be close to actualization. We do not believe that adversarial tampering with ML models is less of a threat today than malicious programs were to early operating systems. The attackers' technology will likely advance, and therefore we need to think about defenses and the possible implications for our policy framework now.

---

53. For example, social media websites compress the images that users upload, which might degrade the adversary's capabilities. Similarly, autonomous vehicles might merge images from different cameras or cut or crop them to suit their model's classification. While preprocessing is highly application-dependent, we point out that it does happen in general to highlight that adversaries may be restricted.

54. *See, e.g.*, APOLLO, http://apollo.auto/ [https://perma.cc/KTC7-RN6A] (last visited Oct. 15, 2019) (demonstrating an example of one open-source implementation of self-driving car technology that employs LiDAR).

55. *See* Liu et al., *supra* note 37.

56. *Id.*

57. *Id.*; A model inversion attack refers to extracting information from a trained model. *See, e.g.*, Matt Fredrikson et al., *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, COMPUTER & COMM. SECURITY (CCS) CONF. (Oct. 12, 2015), https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf [https://perma.cc/XK9S-HC5V].

IV. ANTI-HACKING LAWS

Legislation often reacts to specific threats or harms. The CFAA is a good example.[58] According to popular lore, President Reagan saw the movie *War Games* and met with his national security advisers the next day to discuss America's cyber vulnerabilities.[59] The CFAA is said to be the result of their deliberations. Enacted in 1986, the CFAA aimed to combat computer-related crimes.[60] Since its implementation, the CFAA has been the nation's predominant anti-hacking law. While drafted to combat traditional computer hacking, "the CFAA has evolved into a behemoth of a federal felony statute."[61] This Part lays out the statutory definitions that the CFAA relies on for applicability, e.g., what is a "protected" computer, etc., and contrasts them with the theme throughout the CFAA's actual usage that shows "almost anything with at least a microchip and some relation to interstate commerce is a protected computer and open to CFAA prosecution."[62]

A. CFAA STATUTORY LANGUAGE

The CFAA is designed to disincentivize the compromising of "protected computers" by threat of prosecution or civil lawsuit.[63] It defines a computer as any "electronic, magnetic, optical, electrochemical, or other high speed data processing device performing logical, arithmetic, or storage functions, and includes any data storage facility or communications facility directly related to

58. *See* Obie Okuh, Comment, *When Circuit Breakers Trip: Resetting The CFAA To Combat Rogue Employee Access*, 21 ALB. L.J. SCI. & TECH. 637, 645–46 (2011). While the CFAA is perhaps the best-known anti-hacking statute, it is hardly the only law or standard to address computer security. Similar laws make roughly the same assumptions as the CFAA. For example, at an international level, the Budapest Convention on Cybercrime lists and defines the crime of "illegal access," i.e., "the access to the whole or any part of a computer system without right." Eur. Consult. Ass., ETS 185 Convention on Cybercrime, Budapest, 23.XI.2001.While the Federal Trade Commission (FTC) does not have a stated definition of hacking, it has developed a series of investigations and complaints involving inadequate security. For example, where a company's security practices sufficiently fall short of best practice, the FTC pursues the company under a theory of "unfair or deceptive practice." *See* 15 U.S.C. § 45 (2012). These proceedings invariably involve the exposure of personal information due to inadequate security protocols and envision hacking in the same way as the CFAA. *See* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014).

59. Popular lore seemingly supported by *War Games* is also mentioned in H.R. REP. NO. 98-894, at 3696 (1984).

60. Matthew Ashton, Note, *Debugging The Real World: Robust Criminal Prosecution In The Internet of Things*, 59 ARIZ. L. REV. 805, 813 (2017).

61. *Id.*

62. *Id.*

63. Computer Fraud and Abuse Act, 18 U.S.C.A. § 1030(c)(4)(i)(I)(g) (2008) [hereinafter CFAA].

or operating in conjunction with such device."[64] The CFAA specifically excludes from its ambit "automated typewriters or typesetter, a portable hand-held calculator, or other similar device."[65]

Protected computers are computers

> exclusively for the use of a financial institution or the United States Government, or, in the case of a computer not for such use, used by or for a financial institution or the United States Government and the conduct constituting the offense affects that use by or for the financial institution of the Government.[66]

The CFAA also protects any computer, regardless of its connection to the government, "which is used in or affecting interstate or foreign commerce or communication, including a computer located outside the United States that is used in a manner that affects interstate or foreign commerce or communication of the United States."[67] The courts have deferred to the government on the former definition.[68] The latter definition seemingly encompasses any computer with connections to the United States, but carries with it certain limitations around damages discussed below.

The CFAA applies to both external and internal actors trying to compromise protected computers.[69] External actors incur liability when they "intentionally access a [protected] computer *without authorization*."[70] Internal persons, such as current or former employees, face liability if they already have access to a protected computer, but use the system in such a way that "*exceeds* [their] authorized access."[71] For example, an employee may have access to an email database, but abuse that access by downloading it and sharing the contents with a competitor. Further, an employee may have access to Social Security records, but may exceed her authorized use if she looks up citizens' records without a legitimate business purpose, as was the case in *United States v. Rodriguez*.[72] However, cumulative case law demonstrates that the definition of "internal persons" includes users who persist in violating a terms of service despite being warned against it.

64. *Id.* at § 1030(e)(1).
65. *Id.*
66. *Id.* at § 1030(e)(2)(a).
67. *Id.* at § 1030(e)(2)(b).
68. United States v. Nosal, 676 F.3d 854 (9th Cir. 2012).
69. For example, the CFAA can implicate a rogue employee damaging a company from within *or* it can attach liability to a third-party trying to hack into a system from the outside.
70. CFAA, *supra* note 63, at § 1030(a)(1).
71. *See* United States v. Rodriguez, 628 F.3d 1258 (11th Cir. 2010).
72. *See id.*

What is important for our discussion is that the CFAA prohibits not only "accessing" a computer to "obtain" information, but also "knowingly caus[ing] the transmission of a program, information, code, or command, and as a result of such conduct, intentionally causes damage . . . to a protected computer," as long as this conduct "causes damage *without authorization*."[73] Thus, for example, a code that encrypts a hard drive or corrupts data, or a botnet attack that shuts down a server, can violate the CFAA even though no information has been obtained. Additional ways to violate the CFAA involve espionage, extortion, and trafficking in passwords. However, by the terms of the statute, there exists no liability for the design or manufacture of a hardware or software with vulnerabilities.[74]

The CFAA has both a criminal and civil component.[75] The criminal component is tiered, with penalties as high as twenty years of imprisonment for repeated offenses or offenses that threaten death or bodily injury.[76] The CFAA defines attacking a government computer as a per se violation if the computer is being used "in furtherance of the administration of justice, national defense, or national security."[77] The CFAA's civil cause of action was added in a 1994 amendment and aimed to remedy any persons who suffered damage or loss from a violation of the CFAA.[78] The CFAA's civil cause of action is narrower than its criminal counterpart, with potential offenders triggering liability only if one of the enumerated provisions in § 1030(c)(4)(A)(i) occurs.[79] The notable civil provisions include potential liability for causing at least $5,000 in aggregate damages within a one-year period, potential or actual harm to a broad range of medical equipment, threatening public safety or health, causing physical injury to someone, or damaging ten or more protected computers within a one-year period.[80]

B. INTERPRETATION OF THE CFAA

The CFAA's statutory text leaves much room to hypothesize how these broad definitional parameters apply to facts on the ground. A series of well-publicized cases helps define the range of situations to which CFAA applies.

---

73. CFAA, *supra* note 63, at § 1030(a)(5).
74. *Id.* at § 1030(g).
75. *Id.*
76. *Id.* at § 1030(c)(4)(E).
77. *Id.* at § 1030(c)(4)(A)(i)(V).
78. *See* Shawn E. Tuma, *What Does CFAA Mean and Why Should I Care? A Primer on the Computer Fraud and Abuse Act for Civil Litigators*, 63 S.C. L. REV. 141, 156 (2011).
79. *See* § 1030(c)(4)(A)(i).
80. *Id.*

Before CFAA liability can result, the actor must try to gain, or exceed, access to a "protected computer."[81] The CFAA gives a non-exhaustive list of what can qualify as a protected computer.[82] Subsequent interpretation has shown that "protected computer" is given quite an expansive definition.[83] Courts have deemed that cell phones are considered computers under the CFAA; furthermore, since cell phones are used in interstate commerce or communication, they would also be considered protected computers.[84] In determining that cell phones count as computers, courts looked at the fact that cell phones keep track of the number of incoming and outgoing calls, i.e., "performing logical, arithmetic, or storage functions" under the CFAA.[85] Further, courts emphasized that cell phones use "software" as part of their integral function.[86]

A court analyzed whether information transmitted without authorization needs to be "malicious" to constitute a CFAA violation. The court in *Fink v. Time Warner Cable* found that the CFAA does not require the information transmitted to be malicious for the actor to incur liability.[87] Here, Time Warner Cable remotely accessed its customers' computers to transmit a "reset packet" to prevent undesired functions by way of throttling peer-to-peer file sharing.[88] The reset packet had no malicious purpose, but the unauthorized access and transmission alone were sufficient to violate the CFAA, and met the CFAA's damage requirement when customers claimed the reset packages diminished the services they purchased.[89]

Courts tend to be particularly expansive in their interpretation of the statute when the facts of the case implicate a public safety concern. In *United States v. Mitra*, the court stretched the CFAA's transmission requirement to include sending out a radio signal.[90] The radio signal was used to interfere with a dispatching station's function for the local police department and 911 call center by jamming the signal.[91] This case illustrates that although the CFAA's transmission element requires the transmission of "a program, information,

81. *See id.* at § 1030(a).
82. *Id.*
83. *See* United States v. Kramer, 631 F.3d 900, 902–03 (8th Cir. 2011) (defining a cell phone as a computer).
84. *See id.*
85. *Id.* at 902.
86. *Id.* at 904.
87. *See* Fink v. Time Warner Cable, 810 F. Supp. 2d 633 (S.D.N.Y. 2011).
88. *Id.*
89. *Id.*
90. United States v. Mitra, 405 F.3d 492, 493–96 (7th Cir. 2005).
91. *Id.* at 493.

code, or command" to trigger CFAA liability,[92] the transmission definition expands more liberally when public safety is compromised.

Subsequent courts have furthered the analysis in cases of blocking access to websites by means of denial of service (DDoS) attacks. In dealing with DDoS attacks against websites, courts focused on the "intent to cause damage" provision of the CFAA.[93] In *United States v. Carlson*, Carlson directed thousands of emails at a single email address to try compromising the function of a website.[94] The court found that Carlson violated the CFAA because he was aware of, and motivated by, the potential damage that his actions could cause.[95] In an analogous case, the defendants attempted to disrupt the operations of a business by directing "swarms" of phone and email messages at their respective addresses.[96] The concentrated attacks at the business's personal accounts were methods "that diminish[ed] the plaintiff's ability to use data or a system . . . caus[ing] damage," and thus violated the CFAA.[97] Therefore, courts have broadened the definition of "hacking" by adding the CFAA liability to blocking access to services or platforms.

Hacking under the CFAA has even been defined to include using a website's services in a way that violates the owner's terms of service, as long as the violator has been adequately warned by the website's owner.[98] In *Facebook, Inc. v. Power Ventures, Inc.*, Vachami violated Facebook's Terms of Use Agreement by sending automated messages to Facebook users and subsequently received a cease and desist letter regarding his actions.[99] By continuing to violate the Terms of Use Agreement, the court concluded Vachami knowingly "exceeded authorized access" and violated the CFAA.[100]

When it comes to computers that do not have terms of service assented to by a user, there is no CFAA liability if the user discovers and exploits a system vulnerability, as long as the user did not circumvent any security protocols programmed into the computer.[101] In an interesting unpublished case, *United States v. Kane*, the defendant discovered that an electronic poker machine had

92. *See id.* at 494.

93. *See* United States v. Carlson, 209 F. App'x 181 (3d Cir. 2006).

94. *Id.* at 183.

95. *See id.*

96. *See* Pulte Homes, Inc. v. Laborers' Int'l Union of N. Am., 648 F.3d 295, 299 (6th Cir. 2011).

97. *Id.* at 301.

98. *See* Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1068 (9th Cir. 2016).

99. *Id.* at 1962–64.

100. *See id.*

101. *See* United States v. Kane, No. 2:11-cr-00022-MMD-GWF, 2015 WL 13738589, at *1 (D. Nev. Dec. 16, 2015) (unpublished cases have limited precedential effect).

a flaw in its software that allowed him to push a series of buttons in a particular order to cause the machine to declare him the winner, resulting in a windfall of earnings.[102] The court agreed with the prosecution in deeming the electronic poker machine a "protected computer," but did not extend CFAA liability to the defendant due to his lack of circumvention, or traditional hacking.[103]

Notably, the CFAA does not have an explicit research exception built into the statute.[104] Thus, security researchers attempting to discover potentially dangerous security flaws in protected computers can, in theory, be prosecuted by the full weight of the CFAA. However, the recent decision in *Sandvig v. Sessions* marks an important step towards recognizing a research exception.[105] In *Sandvig*, four professors and a media outlet desired to perform outcome-based audit testing to look for discrimination on real estate websites, and brought a constitutional challenge to clear them of any CFAA liability.[106] Outcome-based audit testing necessarily requires the creation of fake online profiles, thus breaching the websites' terms of service. While *Sandvig* only addressed the CFAA's criminal components, the court emphasized the academic and journalistic motivations and distinguished them from commercial, malicious, or competitive activities.[107] Ultimately, the court ruled that the outcome-based audit testing constituted "merely a practical use of information . . . but it does not constitute an access violation."[108] While an explicit research exception may be added in the future, the current absence of such an exception for research purposes stands in contrast to other federal laws, such as the Digital Millennium Copyright Act (DMCA).[109]

C. APPLYING THE CFAA TO ADVERSARIAL ML

In this final section, we apply the language and interpretation of the CFAA to a specific set of case studies. These case studies are hypothetical, but grounded in actual research. Again, as we have described above, adversarial ML is subject to certain limitations related in part to the research context. Here,

---

102. *Id.*

103. *See id.*

104. *See* Derek E. Bambauer & Oliver Day, *The Hacker's Aegis*, 60 EMORY L.J. 1051, 1105 (2011).

105. *See generally* Sandvig v. Sessions, 315 F. Supp. 3d 1 (D.D.C. 2018).

106. *Id.*

107. *See id.*

108. *Id.* at 27.

109. Digital Millennium Copyright Act, 17 U.S.C. § 1201 (1998). The DMCA, which prohibits circumventions of copyright protections of digital mediums, has an expressly carved-out research exception specifically for encryption research. The DMCA exempts encryption researchers who "circumvent a technological measure for the sole purpose of . . . performing the acts of good faith encryption research." *Id.*

we assume that techniques of adversarial ML can be transferred into real world settings.

We only analyze the following case studies in light of the CFAA's applicability. The authors recognize that the scenarios below could be illegal under different laws, causes of action, and jurisdictions. The CFAA remains the focus of our analysis because of its broad, federal applicability. There is a marked difference between the FBI being able to pursue an adversary for attempting to undermine infrastructure by perturbing road signs to fool driverless cars under a federal criminal statute versus local municipalities having to pursue the same activities as vandalism or reckless endangerment under local law. Moreover, the language of the CFAA is notoriously broad and flexible, such that generating new borderline cases is particularly interesting. Ultimately, however, we are interested in evidencing the broader disconnect between how the law conceives of hacking and this new generation of adversarial ML techniques.

1. Case Studies

Planting adversarial sound commands in ads. A perpetrator of intimate-partner violence buys a local television advertisement in the jurisdiction he suspects his ex-partner now resides. He embeds the ad with an adversarial sound input that no person would recognize as meaningful. The attack causes his ex-partner's Echo, Google, Home, Siri, or other digital personal assistant in range of the TV to publish her location on social media.

Causing a car crash by defacing a stop sign to appear like a speed-limit sign. An engineer employed by a driverless-car company extensively tests the detector used in the cars. She reports to the founder that she has found a way to knowingly deface a stop sign to trick the car into accelerating instead of stopping. The founder suspends operations of his own fleet, but defaces stop signs near his competitor's driverless-car plant. The defaced stop signs cause the founder's competitor's driverless vehicles to get into an accident, resulting in bad publicity.

Shoplifting with anti-surveillance makeup. An individual steals from a grocery store equipped with facial recognition cameras. In order to reduce the likelihood of detection, the individual wears makeup she understands will make her look like an entirely different person to the ML model. However, she looks like herself to other shoppers and to the grocery store's staff.

Poisoning a crowd-sourced credit rating system. A financial startup decides to train a ML model to detect "risky" and "risk averse" behavior in order to assign creditworthiness scores. A component of the model invites Internet users to supply and rate sample behaviors on a scale from risky to risk

averse. A group of teenagers poison the model by supplying thousands of images of skateboarders and rating them all as risk averse. One teenager from the group whose social network page is full of skateboarding pictures secures a loan from the start up and later defaults.

Data inversion across international borders. A European pharmaceutical company trains and releases a companion model with a drug it produces that helps doctors choose the appropriate dosage for patients. The model is trained on European data. But, it is subsequently released to doctors in the United States. An employee in the United States sells access to a marketing company that uses an algorithm to systematically reconstruct the training set, including personal information.

2. Analysis

A case can be made that the CFAA could apply to each of these scenarios. The adversarial sound in the first scenario could constitute the "transmission" of a "command" to a "protected computer," i.e., the victim's phone.[110] Assuming the revelation of the victim's location leads to physical harm, perhaps in the form of violence by the perpetrator, the damage requirement of CFAA has been satisfied. Similarly, by defacing the stop sign, the malicious competitor can be said to have caused the transmission of "information"[111]—from the stop sign to the car—that led to a public safety risk. In both instances, had the attacker broken into the phone or car by exploiting a security vulnerability and altered the firmware or hardware to cause the precise same harm, the CFAA would almost certainly apply.

On the other hand, a perhaps equally strong case could be made that the CFAA does not apply. In neither scenario does the defendant circumvent any security protocols or violate a term of service. The transmission of an adversarial sound seemingly does not cause damage without authorization to a protected computer. Rather, it causes damage to a person through an authorized mechanism—voice control—of a protected computer. With respect to the driverless car scenario, it seems to be a stretch to say that minor changes to the visual world that a sensor may come across constitute the "transmission" of "a program, information, code, or command" on par with a DDoS attack.[112] Regardless, there is again arguably no damage to the detector "without authorization" as required under § 1030(a)(5)(A).

Notably, the same logic of the driverless car scenario arguably applies to the shoplifter who evades facial recognition, at least for purposes of the CFAA.

110. 18 U.S.C. § 1030(a)(5)(A).
111. *Id.*
112. *Id.*

Like the founder who defaces the stop sign to mislead the car's detector, the shoplifter who alters her face to mislead the facial detector has arguably transmitted information purposely to trick the grocery store into misperceiving her so she can steal. Obviously, there are differences. The founder causes physical harm, the shoplifter financial. The founder has no right to alter a stop sign, whereas the shoplifter has a right to apply makeup to her face. But from the CFAA's perspective, the two situations are analogous.

The example of mis-training the credit rating system contains similar ambiguities. From one perspective, the teenagers exploited a flaw in the design of the system in order to embed a trojan horse in the form of a correlation between skateboarding and creditworthiness. Certainly, if the group circumvented a security protocol and changed the valence of skateboarding by hand, their actions would fall within the scope of the CFAA. From another perspective, however, the teens just played by the rules—however misconceived. The state or startup could prosecute or sue them under the CFAA no more than the designer of the flawed poker machine in *Kane* that paid out every time a specific sequence was entered.

The resolution of the final scenario depends, once again, on whether tricking a system into divulging private information should be interpreted the same as hacking into the system to steal that information. Presumably the European pharmaceutical company—beholden to strict EU law[113]—did not design the model anticipating exfiltration of data. But nor did the perpetrator who accessed the model without authorization. He merely queried the model in a surprising way.

V. WHAT IS AT STAKE

To sum up the argument thus far, contemporary law and policy continue to conceive hacking as breaking into or disabling a computer. Devices increasingly leverage ML and potentially other techniques of AI to accomplish a range of tasks. These "smart" systems are not so smart that they cannot be tricked. A burgeoning literature in computer science uncovered various techniques of adversarial ML that, at least in experimental settings, can mislead machines and even force dangerous errors.[114] A comparison between the leading anti-hacking law and adversarial ML reveals ambiguity. It is simply not clear how or when the CFAA applies to "tricking" a robot as opposed to

113. See Eur. Consult. Ass., ETS 185 Convention on Cybercrime, Budapest, 23.XI.2001 (defining the crime of "illegal access," i.e., "the access to the whole or any part of a computer system without right").

114. *See supra* Section III.

"hacking" it. This ambiguity has a number of potentially troubling consequences, which this Part now explores.

A. LINE-DRAWING AND OVERREACH

Our first concern is that line-drawing problems will lead to uncertainty, which in turn could fuel prosecutorial overreach. The CFAA already faces criticism of its hazy boundaries,[115] since both companies and prosecutors have pushed the envelope in arguably problematic ways.[116] A thoughtless application of the CFAA to adversarial ML could exacerbate the problem by grossly expanding its power.

To illustrate, consider again the problems on subtly defacing a stop sign to make it appear like a speed limit sign and subtly altering one's makeup to fool facial recognition. It seems plausible enough that a prosecutor would bring a CFAA violation in the former case, and a court would permit the state to go forward. A judge may intuitively think that providing false inputs to car detectors is analogous to transmitting malicious code or engaging in a DDoS attack. Coupled with the tendency of courts to be more solicitous of the state in CFAA cases involving public safety hazards, we can readily imagine a judge upholding the state's CFAA theory.

Then what about the latter case? How does a court that rules in favor of the state when the defendant tricked a robot car then turn around and decide against it when the defendant changes her appearance to trick an AI-enabled camera in various contexts? The line cannot be that one intervention can cause physical harm and the other cannot. Tricking a car will not always cause physical harm.[117] On the other hand, fooling facial recognition in theory could cause harm, at least in our shoplifter example. Moreover, the CFAA does not require harm to flow from unauthorized access if the protected computer at issue belongs to the government and is used in furtherance of security. Thus, wearing makeup at the airport with the intent not to be recognized by TSA cameras could rise to a CFAA violation, at least in the wake of a precedent

115. *See generally* Philip F. DiSanto, *Blurred Lines of Identity Crimes: Intersection of the First Amendment and Federal Identity Fraud*, 115 COLUM. L. REV. 941 (2014).

116. *See, e.g.*, hiQ Labs v. LinkedIn Corp., 273 F. Supp. 3d 1099 (2017) (providing an example in which LinkedIn unsuccessfully attempted to use the CFAA to block hiQ Labs from using LinkedIn users' public data); United States v. Drew, 259 F.R.D. 449, 456 (C.D. Cal. 2009); Andy Greenberg, *Oops: After Threatening Hacker With 440 Years, Prosecutors Settle For A Misdemeanor*, WIRED (Nov. 26, 2014), https://www.wired.com/2014/11/from-440-years-to-misdemeanor/ [https://perma.cc/QA4R-RYXA].

117. *See, e.g.*, James Bridle, *Autonomous Trap 001* (2017), https://jamesbridle.com/works/autonomous-trap-001 [https://perma.cc/CB5V-3B3F].

holding that defacing a stop sign with the intent that it not be recognized by a driverless car violates the CFAA.[118]

Note that the CFAA punishes not only hacking, but also any *attempted* offense that "would, if completed" cause damage or loss.[119] This attempt provision also aligns oddly with adversarial ML. Automated attempts to locate and exploit vulnerabilities in protected computers clearly constitute attempts for purposes of the CFAA. But what about wearing anti-surveillance makeup all day, in a variety of settings? And does the person who defaces a stop sign "attempt" to attack each and every car that passes, even if a human is at the wheel? These, too, remain open questions.

B. CHILLING RESEARCH

Our second concern flows from the first. If courts interpret the CFAA too broadly in the context of adversarial ML, then researchers may fear running afoul of the CFAA—which has no research exception—when testing real systems for resilience. The case that the CFAA's overreach chills security and other research has already been made repeatedly.[120] Researchers may fear compromising proprietary systems or scraping digital platforms for data, even if they do not have a malicious purpose. The CFAA has a private cause of action and firms may still have an incentive to chill such research to avoid embarrassment. There are safety valves—such as the requirement of harm for private litigants—but the threat of lawsuit alone could suffice to dissuade some research.

Thus, our argument is not one of kind, but of degree. As noted in the Obama White House report on AI,[121] by the AI Now Institute,[122] and by the U.S. Roadmap to Robotics,[123] independent researchers have a critical role in examining AI systems for safety, privacy, bias, and other concerns. The community relies on the ability of impartial individuals within academia, the press, and civil society to test and review new applications and independently

118. Law enforcement may indeed want facial recognition avoidance to constitute a crime. But our intuition is that most would see facial recognition avoidance as a reasonable means by which to preserve privacy and liberty interests, and in any event, of a different order from tricking a vehicle into misperceiving a road sign.

119. 18 U.S.C. §§ 1030(b)–(c).

120. *See generally* Jonathan Mayer, *Cybercrime Litigation*, 164 U. PA. L. REV. 1453 (2016).

121. NAT'L SCI. & TECH. COUNCIL COMM. ON TECH., EXEC. OFFICE OF THE PRESIDENT, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE (2016), https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf [https://perma.cc/836C-3PUG].

122. ALEX CAMPOL ET AL., AI NOW 2017 REPORT 13 (2017).

123. COMPUTING RES. ASS'N, A ROADMAP FOR US ROBOTICS: FROM INTERNET TO ROBOTICS (2016).

report on their performance. Should courts come to expand the CFAA's ambit to include manipulation of AI for testing purposes, the result would be to remove an important avenue of AI accountability.

In an effort to help encourage security testing and research while providing needed legal clarity to researchers, our team explored how the DMCA's exemption rulemaking process provides important guidance.[124] The DMCA directs that every three years, the Librarian of Congress and the Copyright Register engage in a solicitation of recommendations to add *exemptions* to the DMCA's general anti-circumvention provisions.[125] This exemption process incorporates public notices, written comments, and public hearings.[126] During the 2015 review period, the Copyright Register received nearly 40,000 public comments.[127] While the rulemaking editing process can be a time consuming and costly process, the CFAA-affected community could greatly benefit from an opportunity to give input and help from the CFAA to ensure that Congress's legal intent is still relevant and in line with current technological and research realities.

C. INCENTIVE MISALIGNMENT

The first two concerns deal with an interpretation of hacking that is too broad. The last problem deals with the opposite: if adversarial ML is *not* hacking, then do firms that release AI-enabled products and services have any legal obligation to ensure that these systems remain resilient to attack? As alluded to above, the CFAA is not the only anti-hacking law or policy to assume a particular mental model. The FTC also requires products and services to employ reasonable measures against hacking.[128] If hackers can too easily compromise a system, then the FTC can bring—and repeatedly has brought—complaints against the firms that make those systems.[129]

Tricking a robot can sometimes accomplish functionally the same ends as hacking it. Thus, an adversary might "steal" private information by hacking into an AI-enabled system or by reverse-engineering its training data. Similarly, an adversary could temporarily shut down a system through a DDoS attack or by poisoning its training data to make the system suddenly useless in one or more contexts. To the extent that the prospect of an FTC enforcement action

124. *See generally* Maryna Koberidze, *The DMCA Rulemaking Mechanism: Fail Or Safe?*, 11 WASH. J.L. TECH. & ARTS 211 (2015).

125. *See id.*

126. *Id.* at 218.

127. U.S. COPYRIGHT OFF., FISCAL 2015 ANNUAL REPORT (2015).

128. Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 617–18 (2014).

129. *Id.*

incentivizes firms to take basic precautions against attack, we might worry about the failure of the agency to envision susceptibility to adversarial ML. This failure is akin to poor security that would under-incentivize companies to build robust systems.

It is fair to point out a potential tension here. How could we be arguing, on the one hand, that it is dangerous to widen the scope of hacking to encompass adversarial ML when it comes to the threat of prosecution or litigation, but also be arguing that it is dangerous not to when it comes to security standards? But note that the FTC and other bodies are not limited to enforcing security under broad standards such as "unfairness and deception." The FTC could create a separate category of unfairness for inadequate resilience to known adversarial ML techniques, without committing to the idea that tricking is hacking.

VI. CONCLUSION

Computer security is undergoing a paradigm shift, if not a significant evolution. Computer systems continue to be a target for malicious disruption and exfiltration of data. As contemporary applications increasingly leverage ML and other AI techniques to navigate the digital and physical world, these systems present new concerns, as well. Recent research, including by some of the authors, demonstrates how the added benefits of AI also entail novel means of compromising computers. Researchers have shown in experimental settings that ML can be misdirected during both inference and training and that training data can sometimes be reconstructed. In short, robots can be tricked.

Collectively, the prospect of adversarial ML may require the law and policy to undergo a significant evolution of their own. Contemporary anti-hacking and security laws assume hacking to involve breaking into or temporarily incapacitating a computer with code. The misalignment between the early understanding of hacking and today's techniques creates ambiguity as to where and how the laws apply. This ambiguity is dangerous to the extent that it invites prosecutorial overreach, chills research, or leads to underinvestment in hardening measures by firms releasing ML-enabled products and services.

Ultimately, it is up to courts, policymakers, and the industry to come to grips with the prospect of tricking robots. Our role is not to dictate a precise regulatory framework. We do have a few recommendations, however, that follow from our concerns. We recommend clarifying the CFAA to limit prosecutorial discretion. We recommend clarifying the CFAA and related laws to exempt research into AI resilience, so we can continue to test systems for safety, privacy, bias, and other values. Finally, we recommend incentives for

firms to build AI systems more resilient against attacks, perhaps in the form of FTC scrutiny, should a firm release a cyber-physical system that is too easy (whatever that comes to mean) to trick. This is, of course, represents only the beginning of the conversation. We very much look forward to the thoughts of other experts.