CONSTITUTIONAL RIGHTS IN THE MACHINE-LEARNING STATE

Aziz Z. Huq†

A new class of "machine learning" tools is able to make allegedly better predictions and inferences from data than has previously seemed feasible. For the state, machine learning is a powerful and supple device to reveal citizens' hidden beliefs, actions, and expected behaviors. Its deployment to allocate investigative resources, welfare benefits, and coercive penalties to particular individuals, though, can implicate due process, privacy, and equality interests. The substantive doctrinal frameworks and enforcement regimes for those entitlements, however, arose in the context of human action. Neither is tailored to a machine-learning context. This Article offers a start to the larger project of developing a general account of substantive rules and enforcement mechanisms to promote due process, privacy, and equality norms in the machine-learning state. Cataloging notable state and municipal adoptions of machine-learning tools, it considers how existing constitutional norms can be recalibrated (in the case of due process and equality) or retooled (in the case of privacy). It further reexamines the enforcement regime for constitutional interests in light of machine learning's dissemination. Today, constitutional rights are (largely) enforced through discrete, individual legal actions. Machine learning's normative implications arise from systemic design choices. The retail enforcement mechanisms that currently dominate the constitutional remedies context are therefore particularly ill fitting. Instead, a careful mix of ex ante regulation and ex post aggregate litigation, which are necessary complements, is more desirable.

† Frank and Bernice J. Greenberg Professor of Law, University of Chicago Law School. Thanks for David Freeman Engstrom, Sharad Goel, Margot Kaminski, Mark Lemley, Julian Nyarko, Alan Rozenshtein, and Ravi Shroff for many helpful comments and critical conversations. The Frank J. Cicero Fund provided support for this research. All errors are mine.

INTRODUCTION

A deep skepticism of the state lies at the heart of American constitutionalism.[1] Aspiring toward government under the rule of law, American constitutionalism aims to tame the state's risks to individual entitlements even as it enables con-

---

[1] *See* Judith N. Shklar, *The Liberalism of Fear*, *in* LIBERALISM AND THE MORAL LIFE 21, 24–25 (Nancy L. Rosenblum ed., 1989).

tributions to the public good. Technology mediates this trade-off.[2] The state's power to shape the lives of its citizens, whether for good or ill, has always been a function of the instruments at its disposal.[3] Today, one technology transforming how the state acts is a class of computational tools called "machine learning." These instruments derive predictions and inferences in new ways, often exploiting pools of otherwise largely opaque data.[4] Many encounter machine-learning tools first in the marketplace. Facebook, for example, uses them to determine what clickbait tempts best; Amazon uses them to predict what products you'll likely purchase.[5] In state hands, machine-learning tools do more than recommend dietary supplements or fashion accessories. Rather, they can exploit previously low-value data—e.g., administrative records, criminal justice records, or public surveillance footage—to generate startling insight into people's beliefs, actions, and likely behavior.

Consider some present and future implications. Public surveillance cameras typically produce thousands of hours of footage. This is far too much to be examined manually, at least some very particularized starting inquiry. Machine-learning tools can be cheaply trained to analyze large volumes of footage and to recognize faces or patterns of conduct through analyses that take a fraction of the time and effort needed for human inspection.[6] In a different context, new computational tools

---

2 This is a central theme of JAMES C. SCOTT, SEEING LIKE A STATE: HOW CERTAIN SCHEMES TO IMPROVE THE HUMAN CONDITION HAVE FAILED 24 (1998) (exemplifying the "pattern of relations between local knowledge and practices" and "state administrative routines"). *See also* CHARLES S. MAIER, LEVIATHAN 2.0: INVENTING MODERN STATEHOOD 86–93 (2012) (describing the interaction of technological changes during the Industrial Revolution and the European state).

3 Technology is not the only determinant of this liberal dilemma. The range of institutional forms available to the state also matters. Most famously, the historian Stephen Skowronek underscores the move from a state of "courts and parties" to one channeled through national bureaucracies. STEPHEN SKOWRONEK, BUILDING A NEW AMERICAN STATE: THE EXPANSION OF NATIONAL ADMINISTRATIVE CAPACITIES 1877–1920, at 24, 35 (1982).

4 Sendhil Mullainathan & Jann Spiess, *Machine Learning: An Applied Econometric Approach*, 31 J. ECON. PERSP. 87, 88 (2017) (defining machine learning in terms of its capacity for "out of sample" prediction). For further details on machine learning and its functionalities, see *infra* text accompanying notes 34 to 46 (defining machine learning).

5 *See* SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM 233–34 (2019).

6 James Vincent, *Artificial Intelligence Is Going to Supercharge Surveillance*, VERGE (Jan. 23, 2018, 10:54 AM), https://www.theverge.com/2018/1/23/16907238/artificial-intelligence-surveillance-cameras-security [https://perma.cc/YGR4-3GAY]. Machine-learning-driven analysis of video surveillance, though, is not proof against counterstrategies, such as the use of "adversarial patches" on clothing that undermine common inference strategies. *See* Simen Thys, Wiebe Van Ranst & Toon Goedemé, *Fooling Automated Surveillance Cam-*

can be trained to analyze the way in which a person holds and swipes her cellphone so as to uniquely identify a user.[7] Commercial banks are already using such biometric signatures to enable remote account access.[8] Someday soon, state uses of the same functionality will follow.

Such examples may understate the significance of machine learning. The latter is a "powerful and highly generalizable set of capabilities" that "in principle . . . can be applied to the management of *any complex system*, from the steering and guidance of a car to the shaping of public policy."[9] As such, machine learning can generate action-guiding predictions about who should be detained,[10] who should be deported,[11] who should be audited,[12] who should be fired from state offices,[13] who should be deemed in need of state assistance,[14] and even who should be killed.[15] Across these applications, machine learning has the potential to greatly improve on im-

---

eras: Adversarial Patches to Attack Person Detection, 2019 IEEE/CVF CONF. COMPUTER VISION & PATTERN RECOGNITION WORKSHOPS 49, 49–50, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9025518 [https://perma.cc/55TB-VE2F].

7 Claire Reilly, *The Way You Swipe Your Phone Could Be Used to Track You*, CNET (July 31, 2018, 10:45 PM), https://www.cnet.com/news/the-way-you-swipe-your-phone-could-be-used-to-track-you/ [https://perma.cc/3RJU-WVPR].

8 Alison Arthur & Bethany Frank, *Five Examples of Biometrics in Banking*, ALACRITI (May 8, 2019), https://www.alacriti.com/biometrics-in-banking [https://perma.cc/ZS8Z-UEVY].

9 ADAM GREENFIELD, RADICAL TECHNOLOGIES: THE DESIGN OF EVERYDAY LIFE 226 (2017) (emphasis added).

10 Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1072–76 (2019) (describing the use of machine-learning tools in bail and sentencing contexts).

11 Spencer Woodman, *Palantir Provides the Engine for Donald Trump's Deportation Machine*, INTERCEPT (Mar. 2, 2017, 11:18 AM), https://theintercept.com/2017/03/02/palantir-provides-the-engine-for-donald-trumps-deportation-machine/ [https://perma.cc/D2LK-EAYR] (reporting that the Department of Homeland Security (DHS) awarded a private contractor a $41 million contract to build an "Investigative Case Management" system to allow DHS to "access a vast 'ecosystem' of data to facilitate immigration officials in both discovering targets and then creating and administering cases against them").

12 Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1163 (2017).

13 Derek W. Black, *The Constitutional Challenge to Teacher Tenure*, 104 CALIF. L. REV. 75, 92–96 (2016) (describing federally mandated adoption of "value-added models" for teacher evaluation).

14 Colin Lecher, *What Happens When an Algorithm Cuts Your Health Care*, VERGE (Mar. 21, 2018, 9:00 AM), https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy [https://perma.cc/J9RD-3KMJ].

15 Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV. (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ [https://perma.cc/7D94-2FD2] ("The U.S. military is pouring billions

perfect human action or, alternatively, to generate new social costs and also to compound malign forms of social stratification.

This Article documents this ongoing technological shift in state action. I then analyze how important individual rights to due process, equality, and privacy can be conceptualized and then implemented in a context of growing state reliance on machine learning. My first aim is hence descriptive. I highlight a subset of ground-level applications of the machine-learning state that most sharply implicate rights-related concerns. While new computational-tools technology can be used at many different points of the policy-making, legislating, and administrative processes, I think the sharpest normative concerns are likely to arise when an algorithm proximately causes a state benefit or penalty to be assigned (or withheld) to (from) a specific individual.[16] This is not, to be sure, the only application of machine learning that can generate worries. Normative concerns can also arise when a machine-learning tool is used to allocate investigative resources, especially when becoming a target of investigation has immediate costs. Documenting both existing and likely future uses of machine learning, particularly by state and local governments, I draw attention to ways such deployments can implicate due process, equality, and privacy concerns. I do not claim such worries are wholly new. In some instances, constitutional concerns track those presented by human action. At other instances, novel worries arise when a machine is involved.

This descriptive exercise exploits the fact that a disparate scattering of plaintiffs are starting to challenge algorithmic instrument in federal and state court.[17] Cases have arisen in the bail and sentencing context in Wisconsin,[18] California,[19]

---

into projects that will use machine learning to pilot vehicles and aircraft, identify targets, and help analysts sift through huge piles of intelligence data.").

16 Those concerns are not wholly absent where individualized determinations are not at stake, but I will focus here on cases of individualized machine determinations because they present the constitutional issues most acutely.

17 An algorithm is "any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output." THOMAS H. CORMEN, CHARLES E. LEISERSON, RONALD L. RIVEST & CLIFFORD STEIN, INTRODUCTION TO ALGORITHMS 5 (2d ed. 2001) (emphases omitted). Machine learning tools are a distinctive subset of algorithms; most of the algorithms challenged in the cases discussed here have been simpler beasts.

18 State v. Loomis, 881 N.W.2d 749, 753 (Wis. 2016).

19 People v. Superior Court (*Chubbs*), No. B258569, 2015 Cal. App. Unpub. LEXIS 105, at *3 (Cal. Ct. App. Jan. 9, 2015).

Ohio,[20] and New York.[21] Litigation often hinges on whether a particular algorithm can be disclosed consistent with trade secrets law.[22] Legal questions are not confined to the criminal justice realm. In Houston, a teachers' union brought an action against an algorithmic tool used to evaluate job performance and determine discharges on due process grounds.[23] In Arkansas, state disability recipients filed suit against the Arkansas Department of Human Services alleging that an "unlawful switch to the computer algorithm" had violated the state's administrative procedure act.[24] None of these cases, though, grapple head-on with the novel questions presented by constitutional challenges to the machine-learning state. To the contrary, their evasion of these questions hints at a need for more systemic thinking about how relevant constitutional norms should be adapted and how existing regulatory and litigation structures are best retrofitted to achieving constitutional compliance today.

Having established a descriptive baseline, I develop two lines of normative analysis. The first takes up ways in which norms of due process, privacy, and equality might be usefully recalibrated as the state shifts from human to machine action. Second, I offer a general account of how the enforcement regime for these rights might best account for the distinctive qualities of a machine-learning state. I sketch the core points of both analytic arcs in brief here.

In regard to the first question of constitutional substance, I focus on due process, equality, and informational privacy concerns because they seem to be the rights most immediately pertinent in the machine-learning state. Whereas the Court has developed detailed doctrinal accounts of due process and privacy, the constitutional law of informational privacy is thin.

---

20 State v. Jennings, No. 2013 CA 60, 2014 Ohio App. LEXIS 2248, at *13 (Ohio Ct. App. May 30, 2014).

21 Flores v. Stanford, No. 18 CV 2468 (VB), 2019 U.S. Dist. LEXIS 160992, at *11–12 (S.D.N.Y. Sept 20, 2019).

22 *See Chubbs*, 2015 Cal. App. Unpub. LEXIS 105, at *9; *Loomis*, 881 N.W.2d at 761.

23 Hous. Fed'n of Teachers, Local 2415 v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168, 1171 (S.D. Tex. 2017) (challenging "the use of privately developed algorithms to terminate public school teachers for ineffective performance" on due process grounds).

24 Ark. Dep't of Human Servs. v. Ledgerwood, 530 S.W.3d 336, 339 (Ark. 2017); *see also* Michael T. v. Bowling, No. 2:15-CV-09655, 2016 U.S. Dist. LEXIS 123749, at *7–9 (S.D. W. Va. Sept. 13, 2016) (reviewing a due process challenge to algorithmic benefits calculation for the developmentally disabled); K.W. v. Armstrong, 180 F. Supp. 3d 703, 706–07 (D. Idaho 2016) (reviewing a due process challenge to software used to calculate Medicaid benefits).

Even accounting for this difference in the degree of doctrinal development, a gap separates extant doctrinal formulations of all three rights and the technological terrain of machine learning. Present doctrinal formulations do not necessarily track the values underlying rights to due process, equality, or privacy when the focus shifts from human to machine action. (Perhaps those doctrinal formations are a bad match to more mundane nonmachine institutional settings and problems. But making that point is not my concern here). Even if they do not completely displace human judgment, and even if prior dispensations entailed some human reliance upon structured decision-making tools such as checklists or simple algorithms, I contend that machine-learning tools raise constitutional concerns in different ways from human action. Yet constitutional rights have been calibrated with human behavior in mind.[25]

My modest aim here is to suggest in a preliminary way some ways in which doctrine can be adjusted or extended given the novel technological landscape. To emphasize, these are suggestions and not definitive prescriptions. The technological and social landscape is changing rapidly. It would be foolish to aver certainty. I aim here to start a conversation and not to provide conclusive answers.

Technological changes places pressure on the formulation of due process, equality, and privacy interests in subtly different ways. For example, in the most familiar cases that courts have historically addressed, due process is advanced by giving regulated subjects an opportunity of a hearing before an individual adjudicator or an appeal to a new adjudicator. If we are concerned with minimizing the net volume of false positives and false negatives, however, there is reason to believe that a human appeal of a machine decision will often be counterproductive. Rather, due process may require changes to a classifier to reduce the risk of errors.

An example of the constitutional equality implications of changing from human to machine-derived judgments involves the calculation of recidivism risk in the criminal justice system. How should discrimination be defined and policed here? On the one hand, the increasing use of computational prediction tools may well reduce the opportunities for implicit or explicit

25 In addition, because lawyers and judges are not trained in either computer science or statistics, understanding of how machine-learning tools work—and how they are similar to, or diverge from, other governance instruments—is not yet widespread. Obviously, this article is an effort to start filling that gap—albeit from the perspective of a lawyer and not a computer scientist or statistician!

bias on the part of adjudicators such as judges and magistrates to influence decisions. On the other hand, those same tools may also embed assumptions about racial and ethnic groups in ways that reproduce undesirable patterns of residential, economic, and social stratification. This can happen without any intentional discrimination, and can involve a number of quite different mechanisms. Whereas equality-related regulation of human actors might usefully focus on concepts of bias and discriminatory intent, it may be more useful to consider computational predictive tools in terms of their predictable disparate effects.

Finally, consider privacy. Constitutional rules under the Fourth Amendment regulate how the state collects data about its citizens and other regulated subjects and have little to say in how that information is used.[26] A technology that allows the state to exploit publicly available data—surveillance footage, public records, and commercial records not protected by the Fourth Amendment—for insights into individual conduct means the state can eschew surveillance regulated by the Fourth Amendment and yet acquire the same information with relative ease. Thanks to technological change, therefore, the existing Fourth Amendment will increasingly fail to shelter constitutional privacy interests. Indeed, the risk to privacy from the state might soon emerge through quite unexpected vectors, for instance through the incidence of data theft from the databases that the state creates in order to implement machine-learning tools.[27]

There is a second, somewhat more abstract, reason for looking closely at the implementation of constitutional rights in the machine-learning state. Knowledge and understanding of

---

26 For an analysis of technological change's influence on surveillance, see Jack M. Balkin, *The Constitution in the National Surveillance State*, 93 MINN. L. REV. 1, 3 (2008) (describing the "National Surveillance State [as] a special case of the Information State—a state that tries to identify and solve problems of governance through the collection, collation, analysis, and production of information"). In contrast, there is surprisingly little scholarship on how the state *uses* information it can collect without constitutional regulation. For a prescient but lonely treatment of use restrictions under the Fourth Amendment, see generally Harold J. Krent, *Of Diaries and Data Banks: Use Restrictions Under the Fourth Amendment*, 74 TEX. L. REV. 49 (1995) (arguing that the reasonableness of a seizure extends to uses even after law enforcement seizes information).

27 Consider, for example, the risk of data breaches that comes with expanded algorithmic capacity. *See* Owen Daugherty, *Oregon State Agency Suffers Data Breach, Potentially Exposing Personal Information*, HILL (Mar. 21, 2019, 6:20 PM), https://thehill.com/homenews/state-watch/435218-oregon-state-agency-suffers-breach-potentially-exposing-personal-data [https://perma.cc/TBB8-CFQ5]; *see infra* Part II.C (discussing privacy implications of data breaches).

computational tools are presently not widely shared. The general public in particular lacks a clear or precise understanding of those instruments or their limits. Machine learning is taking root in the state even before legal professionals have absorbed all that much technical knowledge or practical understanding. It is reasonable to predict that new adoptions of machine learning will endow the state with new capabilities, but will also be distinctly difficult to understand from the perspective of both participants in the legal system and the public. Indeed, it is plausible to worry that increases in state power will be correlated with a diminishing capacity on the part of regulated subjects to understand or challenge exercises of that power.[28] To be sure, this asymmetrical effect may be buffered by the efforts of well-meaning computer scientists to educate the public and the legal profession about machine learning. But I am skeptical that such efforts will be sufficient. As a result, state adoptions of predictive and inference tools are likely to increase the difficulty that citizens have monitoring and responding to its activities, even as the scope of those activities grows.

The second main contribution of this Article is an analysis of the institutional arrangements through which constitutional values might best be vindicated. At present, constitutional norms of due process, privacy, and equality are in the main developed and vindicated via a common-law process of discrete, incremental, and ex post litigation. The process largely relies on the "liability in tort" model commonly identified with the common law.[29] In previous work, I have criticized the discrete and individuated forms through which constitutional rights are enforced in the ordinary course of nonmachine governance. I have suggested that they too often fail to properly constrain the state and also for embodying controversial and regressive moral intuitions.[30] I have also argued in favor of

---

28 *Cf.* JAMIE SUSSKIND, FUTURE POLITICS: LIVING TOGETHER IN A WORLD TRANSFORMED BY TECH 168–87 (2018) ("The future state, armed with digital technologies, will be able to monitor and control our behaviour much more closely than in the past."). The literature's relative inattention to machine learning and other analytic tools is perhaps a result of the Constitution's direct regulation of information acquisition through the Fourth Amendment and its more diffuse and indirect regulation of information processing and use.

29 Steven Shavell, *Liability for Harm Versus Regulation of Safety*, 13 J. LEGAL STUD. 357, 357 (1984).

30 *See, e.g.*, Aziz Z. Huq & Genevieve Lakier, *Apparent Fault*, 131 HARV. L. REV. 1525, 1547–48 (2018) (arguing that courts require apparent fault (i.e., that a defendant violated not only the law but also a social understanding of legality) before remedying constitutional wrong); Aziz Z. Huq, *Habeas and the Roberts Court*, 81 U. CHI. L. REV. 519, 581–86 (2014) (arguing that habeas review applies a similar fault regime); Aziz Z. Huq, *Judicial Independence and the Rationing of*

conceptualizing constitutional harms in terms of systemic dynamics implicating collective interests.[31]

Consistent with those arguments, I argue here that the constitutional concerns raised by machine-learning tools, like many other public policies, are best addressed through a mix of ex ante regulation and aggregate litigation (i.e., litigation seeking to vindicate the interests of a specific individual). Outside the machine-learning state, this aggregative model has largely failed. This defeat is in large measure due to judges' hostility toward certain constitutional rights (and perhaps also to certain populations, such as criminal defendants and prisoners). But the novelty of computational tools presents an opportunity for doing better. I thus press here the possibility that the machine-learning state is well suited to a combination of ex ante regulation and ex post collective auditing (albeit without assuming that non-algorithmic policies would not benefit from this same approach).

In particular, I explore the application of strategies of ex ante regulation, such as technology mandates and transparency regimes of various forms. One aim of such interventions is to facilitate ex post inquiry into whether and how a machine-learning tool behaves "in the wild" (which may be quite different from how it behaves "in the lab"). Then, in respect to auditing instruments through ex post litigation, I underscore the utility of wholesale, prospective, and system-wide forms of relief. Again, nothing in what follows should be construed to imply that similar mixes of regulation and aggregate litigation would be inapt for other contexts. Quite the contrary. Perhaps the "shock of the new" in the machine-learning context will prompt a more general reconsideration of how we regulate to achieve constitutional rights.

The argument proceeds as follows. Part I recounts how the state leans increasingly on machine-learning tools as aid or substitute for human decision making. Part II considers how due process, privacy, and equality values might be recalibrated. Part III then examines how ex ante regulation and ex post aggregate litigation might be combined to ensure that

---

Constitutional Remedies, 65 DUKE L.J. 1, 70–74 (2015) (noting that a fault regime for constitutional remedies leads to unequal treatment of constitutional wrongs, unequal vindication of constitutional rights, and unequal treatment of litigants).

31 *See, e.g.*, Aziz Z. Huq, *The Consequences of Disparate Policing: Evaluating Stop and Frisk as a Modality of Urban Policing*, 101 MINN. L. REV. 2397, 2438–39 (2017) (arguing that police misconduct fails to breed collective efficacy).

machine-learning instruments remain consistent with constitutional norms.

I
THE MACHINE-LEARNING TURN IN GOVERNANCE

In the last decade, advances in the computational science of machine learning have enabled new functionalities of prediction and inference.[32] The state leverages these new tools to vindicate traditional policy ends or to pursue novel goals. Whatever the consequent hazard to constitutional values, there is little chance that the state will forego these new technologies. Quite apart from their efficiency gains, the United States is under intense pressure from domestic interest groups, such as big tech firms, and from geostrategic competitors to accelerate development and diffusion of machine learning.[33] One reason to analyze constitutionalism in the machine-learning state is thus the political inevitability of the latter's adoption in the context of a growing range of state functionalities. To that end, this Part describes the core of the technology at issue, recent and impending state and local adoptions, and some of the ensuing litigation challenges.

A. New Instruments of Prediction and Inference

In general terms, a machine learning algorithm is a computational tool designed to solve a "learning problem . . . of improving some measure of performance when executing some task, through some type of training experience."[34] At an operational level, machine learning has been described in simple terms as follows: "You give the machine data, a goal and feedback when it's on the right track – and leave it to work out the

---

32 *See* Jonathan Schmidt, Mário R. G. Marques, Silvana Botti & Miguel A. L. Marques, *Recent Advances and Applications of Machine Learning in Solid-State Materials Science*, 5 NPJ COMPUTATIONAL MATERIALS 1, 1–2 (2019).

33 For a political economy of machine learning's adoption by the state, see Mariano-Florentino Cuéllar & Aziz Z. Huq, *Privacy's Political Economy and the State of Machine Learning: An Essay in Honor of Stephen J. Schulhofer*, 72 NYU ANN. SURV. AM. L. 14–18 (forthcoming 2020).

34 M. I. Jordan & T. M. Mitchell, *Machine Learning: Trends, Perspectives, and Prospects*, 349 SCIENCE 255, 255 (2015); *see also* Susan Athey, *The Impact of Machine Learning on Economics*, *in* THE ECONOMICS OF ARTIFICIAL INTELLIGENCE 507, 509 (Ajay Agrawal, Joshua Gans & Avi Goldfarb eds., 2019) ("[M]achine learning is a field that develops algorithms designed to be applied to data sets, with the main areas of focus being prediction (regression), classification, and clustering or grouping tasks.").

best way of achieving the end."[35] The common method of supervised learning,[36] for example, entails first supplying an algorithm with a labeled set of training data[37] and then instructing it to derive (or learn) a rule that discriminates between two subsets within the training sample.[38] Thus, the training data might comprise a set of images, labeled "dog," "cat," and "rat." The algorithm might then be instructed to learn a rule to separate images of dogs from cats or rats. Supervised learning can be binary or multiclass, as in this example.[39] It can also entail estimation of a continuous rather than a categorical variable. Using a random starting formulation of a decision rule, the algorithm will at first do no better than random at predicting the right subset. But by perturbing the rule and evaluating whether changes produce more or less accurate results, the algorithm can "learn" a rule that does predict well how the data's features map onto those subsets.[40] This classifying rule, though, is not the direct result of human design.

Notwithstanding the simplicity of this explanation, machine-learning tools can be highly complex in ways that defeat any effort at either facile explication or reverse engineering. To be sure, there is a real debate about whether machine-learning

---

35 HANNAH FRY, HELLO WORLD: BEING HUMAN IN THE AGE OF ALGORITHMS 11 (2019); *see also* JERRY KAPLAN, ARTIFICIAL INTELLIGENCE 32 (2016) (providing a similar colloquial description).

36 Jordan & Mitchell, *supra* note 34, at 257 (defining supervised learning as a process in which "the training data take the form of a collection of (x, y) pairs and the goal is to produce a prediction y^* in response to a query x^*"). Note that this definition is framed in terms of binary classification. This process can also be described in terms of a "classifier," rather than a function, that examines inputs with "feature values" and outputs a class variable. Pedro Domingos, *A Few Useful Things to Know About Machine Learning*, 55 COMM. ACM 78, 79–80 ("A *classifier* is a system that inputs (typically) a vector of discrete and/or continuous *feature values* and outputs a single discrete value, the *class*."). An *unsupervised* machine-learning algorithm begins with unlabeled training data and develops classifications based on the data's immanent structure. PETER FLACH, MACHINE LEARNING: THE ART AND SCIENCE OF ALGORITHMS THAT MAKE SENSE OF DATA 14–17 (2012).

37 *See* COMM. ON THE ANALYSIS OF MASSIVE DATA ET AL., FRONTIERS IN MASSIVE DATA ANALYSIS 104 (2013).

38 ETHEM ALPAYDIN, MACHINE LEARNING: THE NEW AI 46–47 (2016) ("A *class* is a set of instances that share a common property . . . there exists a formulation of the class in terms of those [certain] characteristics, called a *discriminant*.").

39 *See* Javaid Nabi, *Machine Learning—Multiclass Classification with Imbalanced Dataset*, TOWARDS DATA SCI. (Dec. 22, 2018), https://towardsdatascience.com/machine-learning-multiclass-classification-with-imbalanced-dataset-29f6a177c1a [https://perma.cc/U9N4-9X2F].

40 ARLINDO OLIVEIRA, THE DIGITAL MIND: HOW SCIENCE IS REDEFINING HUMANITY 96–97 (2017) (exploring the inductive character of machine learning).

tools are fundamentally different from the statistical models that have been in widespread use long before computational power allowed the exploitation of big data.[41] However that debate is resolved, at least some applications of machine learning are clearly so quantitively different from earlier statistical techniques that they might as well be different in kind.

To get a sense of this potential for complexity, consider the example of deep-learning networks. The latter are "deep" in the sense of relying on multiple layers of nodes through which inputs are channeled and processed.[42] Important forms of deep learning are recurrent neural nets (RNN), which are used in text recognition and translation tools, and convolutional neural nets (CNN), which are central to machine vision.[43] Both RNNs and CNNs process large volumes of training data (such as millions of images or large bodies of text) each with thousands or millions of features. They exploit networked structures to process this data in ways that their constituent elements could not do on their own. An early and influential deep-learning instrument, designed by Geoffrey Hinton and colleagues, handled data with some sixty million parameters.[44] Deep networks can perform some inference tasks that simple instruments cannot. Today, the ChronoNet CNN can examine photographic images to estimate the date at which they were taken[45] and inspect electroencephalogram images to predict the incidence of epilepsy and other brain disorders.[46]

The design of any machine-learning tool requires a number of judgments that are not mechanically determined by a computational theory or by the logical forms to algorithmic design. Importantly, choices first need to be made about what training

---

41 *See generally* Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, & Daniel G. Goldstein, *Simple Rules for Complex Decisions*, 138 J. ROYAL STAT. SOC'Y 771 (2020) (arguing that complex decision rules often do not perform better simple predictors).

42 Yann LeCun, Yoshua Bengio, & Geoffrey Hinton, *Deep Learning*, 521 NATURE 436, 438 (2015) (defining deep learning).

43 JOHN D. KELLEHER, DEEP LEARNING 160–62, 181–83 (2019).

44 Alex Krizhevsky, Ilya Sutskever, & Geoffrey E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 25, at 5 (Fernando Pereira, Christopher J.C. Burges, Léon Bottou & Kilian Q. Weinberger eds., 2012).

45 Blaise Agüera y Arcas, Margaret Mitchell, & Alexander Todorov, *Physiognomy's New Clothes*, MEDIUM (May 6, 2017), https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a [https://perma.cc/Q8NU-CYM7].

46 Subhrajit Roy, Isabell Kiral-Kornek, & Stefan Harrer, ChronoNet: A Deep Recurrent Neural Network for Abnormal EEG Identification 1 (May 18, 2018) (unpublished manuscript), https://arxiv.org/pdf/1802.00308.pdf [https://perma.cc/BM2F-HUCK].

data will be used.[47] Different selections of training data will yield different predictive models.[48]

In the state-action context, available data will often be a product of historical state practices, such as the management of public benefits or the policing of a particular geographic area or ethnoracial concentration. If such historical practices were flawed or biased, the data thereby produced may also be deficient or misleading in the sense of incorporating biases, blind spots, or unwarranted assumptions. Such gaps or other deficiencies in the data then precipitate for the designer a further question of about whether (and if so how) corrective measures might be taken.[49]

Then, once a set of training data set is in hand, a designer must decide on which attributes, or "features," of the training data to employ in learning a new rule.[50] Should gender, race, or another protected trait, for instance, be among them? What about variables that might closely and predictably correlate with a protected trait, such as residential ZIP code? What if an impermissible classification or its close proxy is necessary to achieve reasonably good algorithmic performance (however that is defined)?

At the same time, the designer needs to decide on an "outcome variable."[51] An algorithm will optimize a function of the outcome variable and the model parameters (together called the cost function) as a way to generate predictions.[52] Several such outcome variables may be available, and yet none may precisely track the underlying matter of policy interest. The designer must then choose among unreliable proxies.[53] Similarly, the designer must decide which algorithmic method (e.g.,

---

47 In a useful article, Lehr and Ohm call this stage "playing with the data." David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 700–01 (2017) (describing feature selection).

48 ALPAYDIN, *supra* note 38, at 71–84; Susan Athey, *Beyond Prediction: Using Big Data for Policy Problems*, 355 SCIENCE 483, 483 (2017) (explaining that machine-learning "programs take as input training data sets and estimate or 'learn' parameters that can be used to make predictions on new data").

49 Lehr & Ohm, *supra* note 47, at 681–83.

50 *Id.* at 700–01.

51 *Id.* at 672–73.

52 *Id.* In a bit more detail, each possible model (given by a set of parameters like the coefficients in a regression equation) corresponds to a set of predictions of the outcome variable. The cost function defines a "cost" or penalty between predictions and the true (observed) outcome, and then the aim is to minimize that cost. For example, in the familiar context of linear regression, one is trying to minimize the sum of least squares.

53 *Id.* at 675.

naïve Bayes, random forests, neural network, etc.) best fits her problem, a choice which requires her inter alia to decide whether to use a relatively straightforward instrument or to select a more complex deep-learning tool.[54] This methodological choice is no simple matter.[55] Insiders describe "a field in constant tribal warfare" between different approaches.[56] Within this field of contestation, the value of increasingly complex instrument design is particularly contested, with some computer scientists warning that the increasing complexity and sophistication of newer predictive tools has not yielded performance gains sufficiently robust to "translate into real advantages in practice" on real-world problems.[57]

This, moreover, is not the full extent of necessary judgments by our designer. Another important challenge in designing machine-learning tools is the problem of "overfitting."[58] This occurs, in effect, when an instrument has been too good at writing a predictor for the training data without accounting for the fact that the latter is merely a noisy sample drawn from the world. Solutions to overfitting require a measure of judgment about how much to constrain the model's learning from the training data.[59]

Moreover, a computational instrument learns "specific contingencies for particular scenarios."[60] It does not grasp underlying concepts. A consequence of this thin form of "understanding" is that tools can be brittle when confronted with

---

54 OLIVEIRA, *supra* note 40, at 110–11. Note that the choice of features and method is often made simultaneously.

55 Indeed, sometimes researchers mislabel the method that they have in fact chosen. For cases of this, see Adrien Jamain & David J. Hand, *Where Are the Large and Difficult Datasets?*, 3 ADVANCES DATA ANALYSIS & CLASSIFICATION 25, 29–31 (2009).

56 Carlos E. Perez, *The Many Tribes of Artificial Intelligence*, MEDIUM (Jan. 12, 2017), https://medium.com/intuitionmachine/the-many-tribes-problem-of-artificial-intelligence-ai-1300faba5b60 [https://perma.cc/52CG-PRYS] (listing symbolists, evolutionists, Bayesians, kernel conservatives, tree huggers, and connectionists among those warring factions).

57 David J. Hand, *Classifier Technology and the Illusion of Progress*, 21 STAT. SCI. 1, 2 (2006).

58 PEDRO DOMINGOS, THE MASTER ALGORITHM: HOW THE QUEST FOR THE ULTIMATE LEARNING MACHINE WILL REMAKE OUR WORLD 71–72 (2015) (describing overfitting and characterizing it as the "central problem" of machine-learning design); *see also* Krizhevsky et al., *supra* note 44, at 6 (describing technical solutions).

59 *See, e.g.*, Mullainathan & Spiess, *supra* note 4, at 91–93 (describing the process of regularization and empirical tuning to mitigate overfitting with decision tree models).

60 Gary Marcus, Deep Learning: A Critical Appraisal 8 (Jan. 2, 2018) (unpublished manuscript), https://arxiv.org/pdf/1801.00631.pdf [https://perma.cc/G6VG-KQSK].

examples outside their training data. There is a risk that the rate of successful prediction will drop rapidly when an instrument is "confronted with scenarios that differ in minor ways from the . . . ones on which the system was trained show that deep reinforcement learning's solutions are often extremely superficial."[61] "[H]idden feedback loops" can emerge after beta testing.[62] Adversarial tactics, such as the strategic deployment of other machine-learning tools, can also induce misclassification.[63] Such vulnerabilities can have nontrivial, even "catastrophic[ ]," consequences.[64] For all these reasons, it is not safe to assume that a machine-learning tool will operate predictably on data drawn from a different distribution from the training data.

B. The Machine-Learning State

Since the eighteenth century, a central component of state building has involved deepening information-gathering capabilities and eroding private efforts to shield the person from the state's gaze.[65] The state has also sought "legible form[s]" in which to record data about individual citizens for easy "reading, processing, and relaying."[66] Machine learning advances these epistemic projects by introducing new means of exploiting data that public authorities have to hand over for other reasons. In private contexts, machine-learning tools are used for tasks such as ranking (Google's and Netflix's algorithms) and classification (credit-scoring tools and spam blockers).[67] The state can employ the same techniques of ranking and clas-

---

61 *Id.*; *see also* Robin Jia & Percy Liang, Adversarial Examples for Evaluating Reading Comprehension Systems 2 (July 23, 2017) (unpublished manuscript), https://arxiv.org/pdf/1707.07328.pdf [https://perma.cc/7EYH-PAWG] (demonstrating that the accuracy of a language recognition CNN can be halved by inserting ungrammatical "junk" into the data).

62 David Sculley et al., *Machine Learning: The High-Interest Credit Card of Technical Debt*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 28, at 3 (Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, & Roman Garnett eds., 2015).

63 Nicolas Papernot et al., *Practical Black-Box Attacks Against Machine Learning*, PROC. 2017 ACM ON ASIA CONF. ON COMPUTER & COMM. SECURITY 506, 510 (2017).

64 Brenden Lake & Marco Baroni, Generalization Without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks 1 (June 6, 2018) (unpublished manuscript), https://arxiv.org/pdf/1711.00350.pdf [https://perma.cc/RTV9-G56X].

65 SCOTT, *supra* note 2, at 89–92.

66 COLIN KOOPMAN, HOW WE BECAME OUR DATA: A GENEALOGY OF THE INFORMATIONAL PERSON 37 (2019).

67 FRY, *supra* note 35, at 8–9; *see also* DOMINGOS, *supra* note 58, at 8 (citing "pattern recognition, statistical modeling, data mining, knowledge discovery, pre-

sification to infer facts about regulated subjects' past behavior or to predict their future actions. These inferences can then be deployed to advance a wide range of policy ends: improving criminal justice; refining education policy (especially teacher hiring and retention decisions); targeting regulatory inspections (such as restaurant health inspections); identifying youth at risk of criminal conduct or involvement; and predicting individual financial outcomes such as default.[68] At the same time, there is no reason why prediction will be used solely for benevolent or wise ends. Predictive instruments are already used overseas to stifle political opposition.[69] As a policy tool, that is, machine learning is not intrinsically "good" or "bad." Its normative valence depends on how it is deployed and what collateral costs it imposes.

This section canvasses current and likely future uses of machine learning by federal, state, and local governments in both civil and criminal domains. In the former, predictive instruments are used to allocate enforcement resources, make employment decisions, and assign benefits. In the latter domain, algorithms are used to direct coercion, in the form of policing resources or incarceration, both before a criminal trial and after sentencing. Adoption of machine learning is, I should emphasize, presently uneven. At the moment, many jurisdictions use predictive instruments that have not been developed using the methods described in subpart I.A. Baltimore, for instance, makes bail decisions using a form generated by the City's Pretrial Release Services containing seven questions and a list of twelve mitigating or aggravating factors.[70] This seems unlikely to endure. A recent study using New York bail data, for example, boasts that deep learning might generate large

---

dictive analytics, data science, adaptive systems, self-organizing systems, and more").

68 Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, & Ziad Obermeyer, *Prediction Policy Problems*, 105 AM. ECON. REV. 491, 494 (2015).

69 *See, e.g.*, Steven Feldstein, *How Artificial Intelligence Is Reshaping Repression*, 30 J. DEMOCRACY 40, 42 (2019) (noting how effective AI technology is for repressing dissent). For a graphic and troubling example, see Paul Mozur, *Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras*, N.Y. TIMES (July 8, 2018), https://www.nytimes.com/2018/07/08/business/china-surveillance-technology.html [https://perma.cc/49SX-2CRB].

70 George Joseph, *Justice by Algorithm*, BLOOMBERG: CITYLAB (Dec. 8, 2016, 12:00 PM), https://www.citylab.com/equity/2016/12/justice-by-algorithm/505514/ [https://perma.cc/87P2-2ZW8]. This comprehensive piece notes both ambiguity in how the instrument was created and how it is applied. "[T]he relationship between risk scores, bail recommendations, and bail decisions remains opaque." *Id.*

efficiency gains in pretrial practice.[71] Given the allure of cost savings (and, no doubt, lobbying by firms wishing to sell predictive instruments and the academics who advise them), states are likely to adopt machine-learning tools over clinical assessments or simple human judgment sooner rather than later. Hence, what follows should be understood as exemplifying, not exhausting, the range of likely near-future uses.

1. *Machine Learning and the Regulatory State*

The use of machine learning to guide enforcement resources, such as restaurant inspectors, tax audits, and fraud detection, is increasingly common.[72] Some instances of these machine-guided discretion raise important ethical and constitutional questions. For example, decisions about how enforcement resources are allocated can raise concerns about racial or ethnic bias.[73] Cases in which a predictive instrument is used to directly assign coercion or benefits to an individual obviously can raise due process worries. And any data aggregate can prompt privacy objections. By way of example, I flag here one machine-learning tool used to allocate investigative resources in a context fraught with normative peril.

This predictive tool was introduced in August 2016 in Allegheny County, Pennsylvania.[74] Allegheny Family Screening Tool (AFST) extracted seventy-one features from a dataset created collaboratively by several state agencies as a basis in order to predict instances of abuse or neglect amongst calls made to a state hotline.[75] An AFST score capturing a risk of abuse was displayed to case workers who receive and screen such calls

---

71 Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, & Sendhil Mullainathan, *Human Decisions and Machine Predictions*, 133 Q.J. ECON. 237, 239–41 (2018).

72 Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 7–8 (2019) (collecting examples); *see also* Katelynn Devinney et al., *Evaluating Twitter for Foodborne Illness Outbreak Detection in New York City*, 10 ONLINE J. PUB. HEALTH INFORMATICS e120, e120 (2018) (reporting on New York's use of Twitter data to guide health inspection of restaurants).

73 Kristen M. Altenburger & Daniel E. Ho, *When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions*, 175 J. INSTITUTIONAL & THEORETICAL ECON. 98, 99 (2019) (finding overreporting for ethnic restaurants).

74 Alexandra Chouldechova, Emily Putnam-Hornstein, Diana Benavides-Prado, Oleksandr Fialko, & Rhema Vaithianathan, *A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions*, 81 PROC. MACHINE LEARNING RES. 134, 138, 143 (2018).

75 *Id.* at 136–38; *see also* VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR 132–42 (2018) (describing AFST's implementation).

and used to inform the decision to investigate or not.[76] An investigation in turn could potentially end in a child's removal from a home. Carefully timed disclosure was meant to avoid excessive reliance on the score at the expense of more granular information.[77] Nevertheless, case workers may presume the AFST score is more accurate than their own observations.[78] Florida implemented a similar predictive tool in 2012,[79] and several states are studying its experience to determine whether to follow suit.[80]

Commentators have raised three normative concerns about the AFST system. First, there is evidence from Allegheny County of racial disparities in the decisions taken with the AFST scores. Black families, for example, appear to experience "disproportionate referrals" based on seemingly innocuous events such as a missed doctor's appointments.[81] The designers of AFST identified a risk that either caseworker animus or correlations between nonracial data (e.g., residential zip code) and race could induce differential treatment of equally at-risk Black and white children.[82]

Second, some observers have raised a concern about the "dehumanizing" effect on parents of having "their entire history . . . summed up in a single number."[83]

Finally, the AFST system draws upon very large stocks of state data by aggregating disparate information. The creation of such aggregates, which might shed considerable light on private facts and behaviors, likely creates a new risk of data breaches.[84] Equality, due process, and privacy, in short, are all potentially in play in this Allegheny County system.

Despite these concerns, the use of machine learning in a form akin to an AFST score appears relatively weakly constrained by constitutional norms. Federal administrative law

---

76 Chouldechova et al., *supra* note 75, at 138–39.

77 *Id.* at 144 (noting that AFST is "a decision-support tool that is presented to call screeners at a specific juncture in the decision-making pipeline").

78 EUBANKS, *supra* note 75, at 141–42.

79 Darian Woods, *Who Will Seize the Child Abuse Prediction Market?*, IMPRINT (May 28, 2015, 10:58 AM), https://imprintnews.org/featured/who-will-seize-the-child-abuse-prediction-market/10861 [https://perma.cc/AEJ8-8H55].

80 Stephanie Cuccaro-Alamin, Regan Foust, Rhema Vaithianathan, & Emily Putnam-Hornstein, *Risk Assessment and Decision Making in Child Protective Services: Predictive Risk Modeling in Context*, 79 CHILD. & YOUTH SERVICES REV. 291, 294 (2017).

81 EUBANKS, *supra* note 75, at 153–54.

82 Chouldechova et al., *supra* note 74, at 141.

83 EUBANKS, *supra* note 75, at 152.

84 *See infra* section II.C. 2–3 (discussing data breaches in more detail).

imposes little check on decisions to forego enforcement[85] or otherwise to manage the "day-to-day" implementation of regulation.[86] Indeed, "nearly unfettered discretion" is "the hallmark of many executive decisions."[87] The deployment of algorithmic technologies may make such evaluation yet more difficult, depending on the nature of the paper record generated by the machine as opposed to the human decision maker. Those against whom enforcement is initiated typically (if not inevitably[88]) will also lack an evidentiary basis to complain about being unfairly singled out on due process or equality grounds. Litigation challenging AFST's equality-related or due process concerns, in short, faces an uphill battle.

2. *Machine Learning and the Allocative State*

Machine-learning tools can be used in the allocation or withdrawal of individualized benefits such as employment or financial aid.[89] In the early 2000s, states such began moving to automate the distribution of public benefit systems in the context of a larger movement to eliminate recipients from welfare.[90] Michigan, for example, introduced an algorithmic tool to detect fraudulent applications for unemployment benefits as part of a larger overhaul of the information technology by the state.[91] Since then, states have increasingly relied on algorithmic tools to allocate both public benefits and state em-

---

85 Heckler v. Chaney, 470 U.S. 821, 832–33 (1985).

86 Norton v. S. Utah Wilderness All., 542 U.S. 55, 64, 66–67 (2004).

87 Mariano-Florentino Cuéllar, *Auditing Executive Discretion*, 82 NOTRE DAME L. REV. 227, 229–30 & n.2 (2006); *accord* Rachel E. Barkow, *Foreword: Overseeing Agency Enforcement*, 84 GEO. WASH. L. REV. 1129, 1130 (2016) ("Most aspects of agency enforcement policy generally escape judicial review.").

88 It is not impossible to imagine complaints about political targeting, such as those levelled against the Internal Revenue Service (perhaps unfairly) from 2014 onward. Alan Rappeport, *In Targeting Political Groups, I.R.S. Crossed Party Lines*, N.Y. TIMES (Oct 5, 2017), https://www.nytimes.com/2017/10/05/us/politics/irs-targeting-tea-party-liberals-democrats.html [https://perma.cc/7DNA-RY36]. Similarly, if a municipality relied on public complaints about restaurants to drive the allocation of enforcement resources, it would also risk potentially biased enforcement patterns. Altenburger & Ho, *supra* note 73, at 101–02.

89 *See* Esther Shein, *The Dangers of Automating Social Programs*, 61 COMM. ACM 17, 17 (2018) (describing machine-learning tools used for Medicaid allocation).

90 EUBANKS, *supra* note 75, at 45–51 (noting that automation resulted in a fifty-four percent increase in denials of food stamps, Medicaid, and cash benefits in Indiana).

91 Robert N. Charette, *Michigan's MiDAS Unemployment System: Algorithm Alchemy Created Lead, Not Gold*, IEEE SPECTRUM (Feb. 16, 2018), https://spectrum.ieee.org/riskfactor/computing/software/michigans-midas-unemployment-system-algorithm-alchemy-that-created-lead-not-gold [https://perma.cc/ZLZ9-T29S].

ployment.[92] Legal challenges to the substitution of algorithm for human decision making in these domains tend to focus on the procedural adequacy of the machine decisions.[93] In particular, plaintiffs underscore the risk of erroneous deprivations. Although less attention is given to equality or privacy concerns, they too may be lurking in the background.

Two examples illustrate how such tools are used and how they are now being challenged in court. A first comes from 2016, when the state of Arkansas adopted an algorithm developed by a company called InterRAI to calculate disability benefits.[94] Its algorithm was not developed using machine-learning methods. Rather, InterRAI is a clinical assessment tool[95] that relies on about sixty "descriptions, symptoms, and ailments" to determine the quanta of home-care provision.[96] (I include it here because it usefully illustrates the kind of challenges that more sophisticated tools might face). According to the suit filed by Legal Aid of Arkansas challenging the InterRAI algorithm on state administrative law grounds, the instrument gave "no weight" to the beneficiary's physician's input.[97] The Supreme Court of Arkansas enjoined the instrument's use on the ground that it had been implemented in violation of the state's administrative procedures act without sufficient notice and comment.[98]

One of the points raised in the litigation was the possibility that the InterRAI tool was brittle in the face of subtle or unusual variations in the way symptoms presented in a particular case.[99] For instance, entering in different evaluations of a person's "foot problems" produced "wildly different scores when the same people were assessed, despite being in the same con-

---

92 Matt Leonard, *Government Leans into Machine Learning*, GCN (Aug. 19, 2018), https://gcn.com/articles/2018/08/17/machine-learning.aspx [https://perma.cc/5JBY-NJQF].

93 *See, e.g.*, Hous. Fed'n of Teachers, Local 2415 v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168, 1176–77 (S.D. Tex. 2017) (arguing that the machine-learning tool used to evaluate, and potentially terminate, teachers violated procedural due process).

94 Lecher, *supra* note 14.

95 It is described in more detail in Brant E. Fries, Lisa R. Shugarman, John N. Morris, Samuel E. Simon & Mary James, *A Screening System for Michigan's Home- and Community-Based Long-Term Care Programs*, 42 GERONTOLOGIST 462, 467 (2002).

96 Lecher, *supra* note 14.

97 Complaint at 9, Ledgerwood v. Ark. Dep't of Human Servs., No. 60CV-17-442 (filed Jan. 26, 2017).

98 Ark. Dep't of Human Servs. v. Ledgerwood, 530 S.W.3d 336, 344–45 (Ark. 2017).

99 Lecher, *supra* note 14.

dition."[100] Similar concerns have been raised for the algorithmic determinations of Medicaid eligibility.[101] In the machine-learning context, the existence of brittleness raises questions about the external validity of the classifier learned from training data.[102]

A second domain in which large pools of government data have been exploited to power algorithmic determinations about specific individuals concerns the hiring and retention of public schoolteachers. Again, this practice is illuminated by recent litigation.

In 2010, the Houston Independent School District moved to "data driven" teacher evaluation.[103] It adopted the Educational Value-Added Assessment System (EVAAS).[104] EVAAS evaluates teachers by comparing their students' average test score gains with statewide average gains to compute a "Teacher Gain Index."[105] A teachers' union, though, persuaded a district court judge that due process was violated when a teacher was fired for a low EVAAS score.[106] It was impossible, the union argued, for teachers to replicate their scores, even with access to the algorithm's underlying code.[107] Yet that score "might be erroneously calculated for any number of reasons."[108] The School District settled the union's suit by disbursing backpay and discontinuing EVAAS's use.[109]

Houston, however, is not alone in reaching for algorithmic solutions in the hiring context. In 2015, the Atlanta Public Schools retained the HireVue company to facilitate teacher hiring.[110] HireVue offers deep-learning tools to extrapolate job performance from facial features and interview performance

---

100 *Id.*

101 *See* Shein, *supra* note 89.

102 *See supra* text accompanying note 59 (discussing technical responses to the problem of overfitting).

103 Hous. Fed'n of Teachers, Local 2415 v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168, 1171 (S.D. Tex. 2017).

104 *Id.* at 1172.

105 *Id.*

106 *See id.* at 1180.

107 *Id.* at 1177.

108 *Id.*

109 Ian Sample, *Computer Says No: Why Making AIs Fair, Accountable and Transparent Is Crucial*, GUARDIAN (Nov. 5, 2017, 7:00 AM), https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial [https://perma.cc/QB68-SNT5].

110 *Atlanta Public Schools + HireVue: Hire A+ Teachers with HireVue Recruitment Software*, HIREVUE, https://www.hirevue.com/customers/atlanta-public-schools-fills-vacancies-teacher-recruitment-software [https://perma.cc/BH9B-6EJ3] (last visited Sept. 25, 2020) [hereinafter *HireVue Hires Teachers*].

from online video interviews.[111] HireVue's materials do not disclose how applicants are evaluated, but their description is consistent with the use of affect-detection software.[112] Nor is it clear whether the Atlanta school district (or other public authorities) is using HireVue's video capture functionality alone, or its suite of predictive tools too.[113]

Other challengers to algorithmic allocations of state benefits have also turned to due process arguments. In Indiana, for example, the automated rejection of a benefit application was successfully challenged in 2012 on the due process ground that the system provided recipients with insufficient information about the deprivations of important welfare benefits.[114] In Michigan, the automated system for flagging fraudulent unemployment benefit applications was challenged on the ground that the system "provide[d] no notice of the allegations brought against them, and that this lack of notice, among other systemic problems, deprives claimants of a fair hearing."[115] In contrast, I have not been able to find examples of challenges to algorithmic allocation systems based on equality or privacy concerns. This may be because due process claims are easier to allege. They require information about how decisions appear to be made and not how different groups experience classification decisions or how data is handled in a back-office context. Alternatively, the gap might be because of the historical origins of procedural due process in challenges to the allocation and withdrawal of welfare benefits.[116] This would mean due process challenges are more readily imagined than equality or pri-

---

111 *Hirevue Video Interview Software*, HIREVUE, https://www.hirevue.com/products/video-interviewing [https://perma.cc/XAT5-L877] (last visited Sept. 25, 2020); Loren Larsen, *HireVue Assessments and Preventing Algorithmic Bias*, HIREVUE (June 21, 2018), https://www.hirevue.com/blog/hirevue-assessments-and-preventing-algorithmic-bias [https://perma.cc/JQB3-UX5Y]. The HireVue site does not disclose what kind of machine-learning tool the company uses. *See HireVue Video Interview Software*, *supra* note 111. But the general description fits the use of deep learning to track elements of facial motion and thereby to create composite scores for various kinds of affect. How this relates to teacher quality is an unexplored question.

112 *See infra* text accompanying notes 194–195 (discussing this possibility).

113 *See HireVue Hires Teachers*, *supra* note 110.

114 Perdue v. Gargano, 964 N.E.2d 825, 832 (Ind. 2012) (finding that "due process requires a more detailed explanation of the reasons underlying an adverse determination").

115 Zynda v. Arwood, 175 F. Supp. 3d 791, 799 (E.D. Mich. 2016).

116 Maggie McKinley, *Petitioning and the Making of the Administrative State*, 127 YALE L.J. 1538, 1624 (2018) ("In a series of cases in the 1970s, litigated largely in the context of public benefits, the Court developed a test for administrative due process . . . .").

vacy ones. It would not necessarily mean that due process problems are more common.

Nevertheless, the racial or privacy effects of benefit distributions may well also be real.[117] To see why, consider work by Khiara Bridges on the intersection of informational privacy and the welfare regime for poor mothers. Bridges's analysis does not concern computational decision tools per se but nonetheless illuminates the possibility of important yet unaddressed normative questions arising from the use of algorithms to allocate public benefits.[118] She underscores the extent to which state aid to poor mothers is conditioned on the disclosure of a good deal of personal information about a mother's behavior and her social context.[119] This deprivation of privacy, Bridges contends, cannot be explained by a concern about the health or well-being of either mother or child. She instead reasons that it "would not even be attempted without the baseline supposition about the group to which she belongs."[120] Bridges's argument resonates with a longer line of sociological and political science work emphasizing how racial stereotypes have tended to shape welfare policy.[121] But the normative concerns she raises may become increasing relevant in the algorithmic context. The public entities that collect information used for algorithmic allocation of benefits, for example, may be more vulnerable to data breaches than private entities such as commercial banks.[122] This would mean that an increasing reliance on those tools for benefit allocations will likely shift more of the

---

117 In May 2019, the Illinois General Legislature passed the Artificial Intelligence Video Interview Act. *See* H.B. 2557, 101st Gen. Assemb. (Ill. 2019), https://www.ilga.gov/legislation/101/HB/PDF/10100HB2557lv.pdf [https://perma.cc/W3SA-SA77]; *IL HB2557*, BILL TRACK 50, https://www.billtrack50.com/BillDetail/1067171 [https://perma.cc/UXB2-SPUE]. The measure, which the governor signed into law in August 2019, imposes notice and consent rules on the use of such tools and also allows applicants to request that their video interviews be destroyed within thirty days of the interview. The act would also limit the sharing of such videos. *See* H.B. 2557.

118 *See* KHIARA M. BRIDGES, THE POVERTY OF PRIVACY RIGHTS 5–6, 8–11 (2017).

119 *Id.* at 1–5.

120 *Id.* at 149.

121 *See* MARTIN GILENS, WHY AMERICANS HATE WELFARE: RACE, MEDIA, AND THE POLITICS OF ANTIPOVERTY POLICY 102 (1999) (describing the racialization of opposition to welfare spending, which has "reflected a preexisting stereotype of blacks as lazy").

122 *See* Danielle Keats Citron, *A Poor Mother's Right to Privacy: A Review*, 98 B.U. L. REV. 1139, 1147 (2018) ("A common source of data breaches involves public hospitals where the personal data of poor mothers is collected and stored.").

burden of data-breach risk to the indigent.[123] Machine learning's adoption would then have a regressive and racially disparate economic impact as well as imposing a burden upon privacy rights.

3. *Machine Learning and the Punitive State: Facial Recognition as a Case Study*

The third domain in which machine learning is increasingly used involves the provision of public security through policing, incarceration, and (in the most extreme cases) force. There is already a large body of literature on how machine learning is deployed in policing,[124] bail and arraignment proceedings,[125] and sentencing.[126] This literature depicts how machine-learning tools and other algorithms are used to generate predictions of future violence or criminality. Location-based predictions, such as those generated by policing applications like PredPol, are used to allocate investigative resources.[127] Other predictions can focus on specific individuals. The COMPAS algorithm, for instance, is used in many jurisdictions to facilitate bail determinations by generating a risk score from one to ten for defendants, a score that

---

123 To be clear, this theory has not been tested empirically; I raise it here as a possibility to be evaluated through regulation or litigation.

124 *See, e.g.*, Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327, 383–85 (2015) (providing a careful catalogue of predictive policing tools); Andrew Guthrie Ferguson, *Policing Predictive Policing*, 94 WASH. U. L. REV. 1109, 1120–44 (2017) (similar); *see also* Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871, 929 (2016) (developing a "framework" for integrating machine-learning technologies into Fourth Amendment analysis).

125 *See* Richard A. Berk, Susan B. Sorenson & Geoffrey Barnes, *Forecasting Domestic Violence: A Machine Learning Approach To Help Inform Arraignment Decisions*, 13 J. EMPIRICAL LEGAL STUD. 94, 110 (2016) (reporting experimental results suggesting gains from machine learning prediction of violence risk); Richard Berk & Jordan Hyatt, *Machine Learning Forecasts of Risk to Inform Sentencing Decisions*, 27 FED. SENT'G REP. 222, 223 (2015) (explaining advantages of machine-learning tools); Richard F. Lowden, *Risk Assessment Algorithms: The Answer to an Inequitable Bail System?*, 19 N.C. J.L. & TECH. ONLINE 221, 230–31 (2018) (listing jurisdictions that have adopted algorithmic tools).

126 Richard Berk, *An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism*, 13 J. EXPERIMENTAL CRIMINOLOGY 193, 195 (2017) (discussing the 2010 decision of the Pennsylvania Board of Probation and Parole to use a machine-learning protocol to generate forecasts of recidivism). *See generally* John Monahan & Jennifer L. Skeem, *Risk Assessment in Criminal Sentencing*, 12 ANN. REV. CLINICAL PSYCHOL. 489, 493–95 (2016) (describing the general context of risk assessment in sentencing).

127 Aaron Shapiro, *Reform Predictive Policing*, 541 NATURE 458, 459 (2017).

provides guidance to a magistrate judge charged with setting or denying bail.[128]

Rather than retreading details of predictive policing and bail algorithms that have been well covered elsewhere, I focus here on a new frontier in the law enforcement deployment of machine learning. This is use of facial recognition technologies to identify individuals from public surveillance and body-camera footage.[129] Facial recognition technologies provide a useful case study of the complex and unpredictable ways that norms of procedural fairness, equality, and privacy interact when the state deploys machine-learning tools to draw inferences from otherwise unilluminating data.

Roughly half of all American adults are already profiled in one or another American law enforcement agencies' facial-recognition database.[130] These can be used to match with visual evidence in specific cases and make arrests.[131] More controversially, they can be used to identify participants of protests against government policies.[132] The rate of its adoption is uncertain. In May 2018, Axon—one of the largest manufacturers of body-worn cameras in the United States—secured a patent on real-time identification of faces caught on an officer's body-worn camera.[133] Then in June 2019, the company announced

---

128 EQUIVANT, PRACTITIONER'S GUIDE TO COMPAS CORE 1–2, 8 (2019), https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf [https://perma.cc/PN7Q-59G5]; *see also In re* Hawthorne v. Stanford, 22 N.Y.S.3d 640, 641–42 (N.Y. App. Div. 2016) (describing the COMPAS assessment tool).

129 Dakin Andone, *Police Used Facial Recognition to Identify the Capital Gazette Shooter. Here's How It Works*, CNN (June 29, 2018, 6:22 PM), https://www.cnn.com/2018/06/29/us/facial-recognition-technology-law-enforcement/index.html [https://perma.cc/HL2R-487J].

130 CLARE GARVIE, ALVARO M. BEDOYA, & JONATHAN FRANKLE, GEORGETOWN LAW CTR. ON PRIVACY & TECH., THE PERPETUAL LINE-UP: UNREGULATED POLICE FACE RECOGNITION IN AMERICA 1 (2016), https://www.perpetuallineup.org/sites/default/files/2016-12/The%20Perpetual%20Line-Up%20-%20Center%20on%C20Privacy%C20and%C20Technology%20at%20Georgetown%C20Law%20-%20121616.pdf [https://perma.cc/EK75-6HRG].

131 *See, e.g.*, State v. Alvarez, No. A-5587-13T2, 2015 N.J. Super. Unpub. LEXIS 1024, at *1–2 (N.J. Super. Ct. App. Div. May 4, 2015) (searching every image in the state's repository to determine if individuals were maintaining more than one identification document).

132 *See, e.g.*, Kevin Rector & Alison Knezevich, *Maryland's Use of Facial Recognition Software Questioned by Researchers, Civil Liberties Advocates*, BALT. SUN (Oct. 18, 2016, 12:01 AM), https://www.baltimoresun.com/news/crime/bs-md-facial-recognition-20161017-story.html [https://perma.cc/U5ER-SC3E] (noting that Maryland's image repository was used to monitor protestors during Baltimore protests).

133 Alex Pasternack, *Body Camera Maker Will Let Cops Live-Stream Their Encounters*, FAST COMPANY (Oct. 8, 2018), https://www.fastcompany.com/

that it was not installing the tool because of reliability concerns.[134] For now, individualized facial-recognition results may not reach officers at a time and in a manner that permits them to act upon the data. But this equilibrium is unlikely to hold.

Facial recognition raises interrelated privacy, procedural fairness, and equality concerns. Consider a much-publicized 2015 study using eight facial traits to identify specific persons.[135] Finding no duplicates among a sample of 3,982 facial images provided by the U.S. Army, it favorably compared the accuracy of facial recognition to that of DNA matching.[136] A 2019 paper, however, observed that this result rested on untested assumption about the statistical distribution of certain parameter values for those traits.[137] It doubted the external validity of the 2015 study. That research, for instance, assumed that human faces are static and unchanging over time. But "ageing, illness, tiredness, the expressions we're pulling or how our faces are distorted by a camera angle" all can alter the values of the eight facial traits.[138] Even if facial recognition were accurate under ideal conditions, police deploy it under nonideal conditions. They indeed use it in rather creative ways. Hence, in New York City, when officers had a partial surveillance shot of a face from a pharmacy larceny, they used a high-quality video image of the actor Woody Harrelson to find matches on the theory that the partial image from the surveillance video looked like Harrelson.[139]

Patterns of error rates in lab-based facial recognition systems are also uneven across racial, gender, and age lines. This is a consequence of using predominantly older, more male and whiter exemplars in training data. One 2018 study of two com-

---

90247228/axon-new-body-cameras-will-live-stream-police-encounters [https://perma.cc/845M-D4KG].

134 Charlie Warzel, *A Major Police Body Cam Company Just Banned Facial Recognition*, N.Y. TIMES (June 27, 2019), https://www.nytimes.com/2019/06/27/opinion/police-cam-facial-recognition.html [https://perma.cc/VNP2-YGR4].

135 Teghan Lucas & Maciej Henneberg, *Are Human Faces Unique? A Metric Approach to Finding Single Individuals Without Duplicates in Large Samples*, 257 FORENSIC SCI. INT'L 514.e1, 514.e5 (2015).

136 *Id.* at 514.e2, e6.

137 Ronald Meester, Bart Preneel, & Sylvia Wenmackers, *Reply to Lucas & Henneberg: Are Human Faces Unique?*, 297 FORENSIC SCI. INT'L 217, 218–20 (2019).

138 FRY, *supra* note 35, at 163.

139 Clare Garvie, *Garbage In, Garbage Out*, GEO. L. CTR. ON PRIVACY & TECH. (May 16, 2019), https://www.flawedfacedata.com/ [https://perma.cc/242U-YEP9]. It is not clear, however, how common use misuses are; anecdotal data is a risky basis of a judgment as to whether there is a real problem.

mercially available facial-recognition tools, IJB-A and Adience, for example, found that both were trained on predominantly white subjects and had errors rates for Black women that were 34.7 percent higher than for white men.[140] In respect to privacy, there is little regulation under federal or state law of the inferences police draw from facial images. There is also some evidence that facial images allow for "category-jumping" inferences about health. For instance, they may enable predictions of postpartum depression from expectant mothers' prenatal image postings.[141]

Facial recognition can also be misused in stark ways. A 2016 study by two Chinese researchers used a training set of 1,856 photos of Chinese men to construct a predictive tool to distinguish two "manifolds" of "criminal" and "non-criminal" face types.[142] Their result was extensively criticized. Their small sample of training data, for example, made overfitting difficult to avoid.[143] Many of their noncriminal faces (but none of the criminal faces) wore white collared shirts, introducing a likely confound. Still, it is not far-fetched to envisage police forces generating "criminal type" lists now based on such uses of facial recognition tools—much as they have tried to use social networks (unavailingly) to create "strategic subject lists" of likely future criminals.[144] However flawed the resulting inferences, they might nonetheless sharply inflect police practice.

---

140 Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACHINE LEARNING RES. 1, 3, 11–12 (2018); *see also* Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 680 (2017) ("[A]lgorithms that include some type of machine learning can lead to discriminatory results if the algorithms are trained on historical examples that reflect past prejudice or implicit bias . . . ."); Kate Crawford, *Artificial Intelligence's White Guy Problem*, N.Y. TIMES (June 25, 2016), https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html [https://perma.cc/AHC5-Q9WT] (noting that sexism, racism, and other forms of discrimination are often built into machine learning). Note that the distributive effects of differential error rates depend on the ratio of false positives and false negatives.

141 Eric Horvitz & Deirdre Mulligan, *Data, Privacy, and the Greater Good*, 349 SCIENCE 253, 253 (2015).

142 Xiaolin Wu & Xi Zhang, Automated Inference on Criminality Using Face Images 2–3 (Nov. 13, 2016) (unpublished manuscript), https://arxiv.org/pdf/1611.04135v1.pdf [https://perma.cc/L3KQ-ZX3B].

143 Agüera y Arcas et al., *supra* note 45.

144 *See* Jeremy Gorner, *Chicago Police Use 'Heat List' as Strategy to Prevent Violence*, CHI. TRIB. (Aug. 21, 2013), http://articles.chicagotribune.com/2013-08-21/news/ct-met-heat-list-20130821_1_chicago-police-commander-andrew-papachristos-heat-list [https://perma.cc/DE7Y-Q4LX]. The Chicago "heat list," however, proved to have little or no predictive value. Jessica Saunders, Priscillia Hunt, & John S. Hollywood, *Predictions Put into Practice: A Quasi-Experimental*

There is little litigation testing the constitutional constraints on algorithmic decision making in the criminal justice context.[145] The case law that does exist focuses on due process questions, touches briefly on equality concerns, and largely ignores privacy values.[146] One reason for this is the absence of effective vehicles for raising legal challenges to machine-learning instruments in the criminal justice context. When it comes to policing, for example, it would be difficult for an individual litigant to challenge the use of a machine-learning tool to allocate policing resources so long the legal basis for his or her encounter with the police was constitutionally sufficient.[147] In addition, systemic challenges filed as class actions to the allocation of policing resources over different geographic spaces are exceedingly rare.[148] Costly to investigate and litigate, they are likely to founder on questions of Article III standing and amenability to Rule 23 class-based resolution.

Some cases have arisen in the context of individualized risk evaluations in pretrial and sentencing. In 2016, for example, the Wisconsin Supreme Court rejected a due process chal-

---

Evaluation of Chicago's Predictive Policing Pilot, 12 J. EXPERIMENTAL CRIMINOLOGY 347, 363 (2016).

145 One reason may be the successful exercise of trade secrets objections by the commercial manufacturers of algorithms. Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1349–53 (2018) (arguing that such trade secrets invocations pose a real problem and contending that new transparency mechanisms are required). Many commonly used machine-learning tools are, in fact, quite simple to program in a common language such as R. *See, e.g.*, *Random Forests*, UC BUS. ANALYTICS R PROGRAMMING GUIDE, https://uc-r.github.io/random_forests [https://perma.cc/5J4Q-G9F5] (last visited Sept. 25, 2020) (providing an introduction to random forests using R). Claims to the effect that the basic method (be it random forests, naïve Bayes, or even a neutral net) is somehow bespoke and hence worthy of trade secrets protection are probably bunk. What is more distinctive is the manner of regularization and empirical testing used to tweak the rule learned by the algorithm to avoid overfitting or achieve other ends. For example, a model might be adjusted to avoid predictions that correlate too closely with race or gender.

146 On the possibility of a Fifth Amendment challenge to interviews designed to elicit information from a defendant for the purpose of assigning him or her an algorithmic classification, see Cassie Deskus, Note, *Fifth Amendment Limitations on Criminal Algorithmic Decision-Making*, 21 N.Y.U. J. LEGIS. & PUB. POL'Y 237, 259–66 (2018).

147 Under Fourth Amendment doctrine, the availability of a legal justification for a police stop obviates any argument that it should be treated as unlawful because of the actual causes of or justifications for the stop. *See* Whren v. United States, 517 U.S. 806, 813 (1996) (rejecting "any argument that the constitutional reasonableness of traffic stops depends on the actual motivations of the individual officers involved").

148 For a rare exception, see Cent. Austin Neighborhood Ass'n v. City of Chicago, 1 N.E.3d 976, 978, 984 (Ill. App. Ct. 2013) (challenging the failure to provide resources to minority neighborhoods in Chicago).

lenge to the COMPAS algorithm based on the defendant's limited ability to challenge the algorithm in broad and general terms, rather than being able to scrutinize the individualized data upon which the algorithm relied in a specific instance.[149] The Court reasoned that the algorithm relied on publicly available data alone. It observed that the defendant could have denied or explained any information used to craft his prediction.[150] In passing, the Court noted that traits such as gender were among the large set of inputs to the defendant's sentence.[151] On their own, the Court cautioned, such factors "may not be considered as the determinative factor in deciding whether the offender can be supervised safely and effectively in the community" consistent with due process.[152]

While lawsuits challenging the use of facial recognition have not yet been lodged, regulatory responses have been set in motion. In May 2019, the San Francisco Board of Supervisors voted to prohibit police adoption or implementation of facial recognition technologies.[153] A raft of other cities, including New York, Las Vegas, Detroit, Boston, and Orlando, have nevertheless embraced the technology. They show no sign of willingness to abandon it. New York City has enacted an ordinance creating an expert board to monitor and make recommendations about how algorithmic technologies are to be deployed.[154] It remains to be seen how such a body would operate and whether it will be able to take on a powerful interest group such as the police.

The Wisconsin decision, like the Arkansas challenge to disability allocation algorithms and the Houston challenge to teacher evaluations, turned almost exclusively on procedural concerns. Yet even as a contentious literature has emerged

---

149 State v. Loomis, 881 N.W.2d 749, 761–62 (Wis. 2016).

150 *Id.*

151 *Id.* at 765 ("[T]he due process implications compel us to caution circuit courts that because COMPAS risk assessment scores are based on group data, they are able to identify groups of high-risk offenders—not a particular high-risk individual.").

152 *Id.* at 760.

153 Kate Conger, Richard Fausset & Serge F. Kovaleski, *San Francisco Bans Facial Recognition Technology*, N.Y. TIMES (May, 14, 2019), https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html [https://perma.cc/74JD-4KZX].

154 Local Law No. 49, N.Y. City Council (N.Y. 2018), https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0 [https://perma.cc/D68B-JMMG] (creating a task force charged with investigating "agency automated decision systems").

analyzing the role of race in the COMPAS algorithm,[155] to date there has been no litigation explicitly challenging those effects. Similarly, there is a dearth of academic or judicial treatment of the privacy-related risks from the creation of large aggregates of data for public security purposes. Still, even if police forces have more resources at their disposal than (say) public hospitals, there is no reason to think that they will be inured to the risk of data breaches.

***

Machine-learning tools are rapidly diffusing across both civil and criminal regulatory domains. They are at the moment sporadically regulated. They consistently raise, however, a common cluster of procedural due process, equality, and privacy concerns. Courts and commentators have glimpsed these concerns. But judges to date have neither offered a coherent account of how they are interlaced nor of how they can be identified, let alone mitigated.

II
APPLYING CONSTITUTIONAL VALUES IN THE MACHINE-LEARNING STATE

Given the rapid and ongoing adoption of machine-learning technologies by federal and state authorities, how should constitutional interests be recalibrated to fit the new terrain fashioned by the machine-learning state? This Part focuses on due process, equality, and privacy values, three constitutional norms repeatedly implicated in the design and operation of predictive tools. It analyzes difficulties that arise in their application to the machine-learning state.

A. Procedural Due Process

A common complaint lodged in court against machine-learning instruments is their failure to give regulated subjects procedural due process.[156] Anecdotal accounts abound of in-

155 *See* Huq, *supra* note 10, at 1047–57 (discussing different definitions of racial disparities in algorithmic classification and suggesting why a definition focused on the potential for stratifying effects is most desirable). For a different analysis, albeit one that is critical of COMPAS in a different way, see Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/Z9JF-LCFY].

156 Due process concerns are central in several cases. *See, e.g.*, Hous. Fed'n of Teachers, Local 2415 v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168, 1171 (S.D. Tex. 2017) (arguing that the teacher evaluation algorithm deprived teachers of

dividuals who have been wrongly classified by an algorithm when the error could have been quickly and easily fixed by human attention. In an influential treatment, for example, data scientist Cathy O'Neill describes an applicant for a welfare benefit who fails an automated, "web-crawling[,] data-gathering" background check.[157] It is only when "one conscientious human being" took the trouble to look into the quality of this machine result that error was discovered and corrected.[158] The implication is that machines are prone to error and that a hearing of sorts before a human adjudicator is a necessary adjustment to any algorithmically driven process.

A granular focus on error in the isolated case, however, is an untrustworthy vehicle for the purposes of due process analysis. I shall argue instead that due process is violated when an algorithm fails to achieve an adequate level of accuracy across the population of regulated cases. Due process concerns hence arise from the calibration of design margins in ways that make relevant errors more rather than less likely. A constitutional analysis must therefore focus upon algorithmic design choices remote in time from the instant in which a human is subject to algorithmic classification. Remedies for a due process deficit are unlikely to take the form of additional human review but rather better algorithmic design. I identify a number of relevant design margins in this spirit. I also emphasize that it is not always possible to eliminate equally false negatives and false positives. A choice, rather, must be made about which to endure. Due process in the algorithmic context—where it is possible to precisely specify ex ante the balance and kind of errors—thus entails normative judgments about the relative cost of different sorts of errors. Although those judgments are implicitly embedded in human decision-making process, they can be isolated and addressed with greater precision in the machine-decision context.

---

due process protections against substantively unfair deprivations of property); State v. Loomis, 881 N.W.2d 749, 760 (Wis. 2016) (arguing that COMPAS risk assessment violated the defendant's right to be sentenced based on accurate information). The challenge to Arkansas's automated disability determinations sounds in state administrative law but relied on a notice concern familiar to due process jurisprudence. Ark. Dep't of Human Servs. v. Ledgerwood, 530 S.W.3d 336, 344–45 (Ark. 2017).

157 CATHY O'NEILL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 152–53 (2016).

158 *Id.* at 153.

1. *Procedural Due Process Norms*

The doctrinally dominant model of procedural due process is narrowly "utilitarian" in its focus on "attaining the most accurate conclusion in the most efficient manner."[159] "Accuracy," in the due process context, is understood to mean a correlation between a decision procedure's outcomes and some empirical ground truth.[160] Alternative conceptions hinging on dignity and the intrinsic value of participation have not gained doctrinal purchase.[161] This instrumental, accuracy-focused account of due process crystallized in the famous three-part test announced in *Mathews v. Eldridge*.[162] The Court here directed attention to "the private interest," the estimated "risk of an erroneous deprivation of such interest" combined with "the probable value . . . of additional or substitute procedural safeguards," and "the Government's interest, including the function involved and the fiscal and administrative burdens that the additional or substitute procedural requirement would entail."[163] These factors are properly considered by looking at adjudicative mechanisms as a whole rather than at the specifics of a single case. In this sense, due process challenges commonly have the flavor of a facial challenge.

In application, the *Mathews* test relies on difficult, perhaps irremediably hard, counterfactual questions about the state's election between potential alternative institutional arrangements, private individuals' behavior under alternative adjudicatory arrangements, and the expected gains to accuracy from

---

159 Martin H. Redish, *Discovery Cost Allocation, Due Process, and the Constitution's Role in Civil Litigation*, 71 VAND. L. REV. 1847, 1863–64 (2018).

160 An alternative conception of accuracy would focus on the expression of confidence (uncertainty) in classifications. *See* Robert J. MacCoun, *The Epistemic Contract: Fostering an Appropriate Level of Public Trust in Experts*, *in* MOTIVATING COOPERATION AND COMPLIANCE WITH AUTHORITY 191, 201–02 (Brian H. Bornstein & Alan J. Tomkins eds., 2015). Although I do not purpose MacCoun's proposal at length, I do later explain how uncertainty and accuracy interact in a functionally important way. *See infra* text accompanying note 185 (discussing the bias/variance trade-off).

161 The dignity rationale is vigorously defended in scholarship. *See, e.g.*, Frank I. Michelman, *Formal and Associational Aims in Procedural Due Process*, *in* DUE PROCESS: NOMOS XVIII, at 126, 127–28 (J. Roland Pennock & John W. Chapman eds., 1977) (underscoring participation values as an element of due process); Jerry L. Mashaw, *Administrative Due Process: The Quest for a Dignitary Theory*, 61 B.U. L. REV. 885, 899 (1981) (advancing a dignitary theory of due process); Martin H. Redish & Lawrence C. Marshall, *Adjudicatory Independence and the Values of Procedural Due Process*, 95 YALE L.J. 455, 504 (1986) (advocating for an independent adjudicator to protect procedural due process).

162 424 U.S. 319, 335 (1976).

163 *Id.*

marginal changes to those arrangements.[164] Its categorical exclusion of noninstrumental considerations from due process analysis has also been controversial. But the test has remained good law for almost fifty years. It can logically be applied in new contexts, including those where machine learning is in use. Indeed, I will suggest that the holistic *Mathews* test may well be easier to apply in the latter context than in many of the institutional domains in which it has previously been wielded.

2. *Application to Machine Learning*

Scholarship concerned with the procedural quality of algorithmic decision making have read *Mathews* to demand that specific notice be given to regulated subjects and that an individualized determination, often involving a human adjudicator, be available. In an early analysis, Danielle Keats Citron argued that constitutionally adequate notice is supplied by an audit trail documenting all "decisions made in a case" and "the actual rules applied in every mini-decision that the system makes."[165] Developing the idea of a hearing right against machine decisions, Citron focused on scenarios in which a human adjudicator is supplied with an algorithmic recommendation and recommended that "agencies should require hearing officers to explain, in detail, their reliance on an automated system's decision."[166] This assumes the availability of human intervention after an instrument has been applied to a specific case.

In a similar vein, Kate Crawford and Jason Schultz have pressed for "procedural data due process [to] regulate the fairness of Big Data's analytical processes with regard to how they use personal data (or metadata . . .)."[167] Like Citron, they seem to conceptualize the entailment of due process in granular, individualistic terms. Notice, in their view, entails disclosure of the "type of predictions" and "the general sources of data" used in the algorithm.[168] They too would require a hear-

---

164 Jerry L. Mashaw, *The Supreme Court's Due Process Calculus for Administrative Adjudication in* Mathews v. Eldridge*: Three Factors in Search of a Theory of Value*, 44 U. CHI. L. REV. 28, 46–51 (1976) (offering these critiques in a somewhat looser formulation).

165 Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1305 (2008).

166 *Id.* at 1307; *cf. id.* at 1284 (rejecting the idea that due process would require access to source code).

167 Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward A Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 109 (2014).

168 *Id.* at 125.

ing, in which an affected person could examine the "data input and the algorithmic logic applied," and then appeal to a "neutral data arbiter" (presumably, a human rather than another machine) to resolve disputes about the quality of analysis and prediction.[169] It is not clear whether Crawford and Schultz think that due process also requires disclosure of (1) the data used in the training and generation of the learned rule or (2) the data about the regulated subject used to make a prediction or classification.

In a somewhat similar vein, Cary Coglianese and David Lehr explicated notice by recommending that individuals receive information "collected about them" and "information about how accurate the algorithm is across individuals when evaluated in a test data set."[170]

The focus of these proposals upon a human appeal of individual cases may, however, miss the best way to vindicate due process interests for a number of reasons. First, as David Lehr and Paul Ohm explain, there are "many ways in which data can be selected and shaped—say, during data cleaning or model training"—that undermine the quality of predictions.[171] Deviations from a tolerably accurate pattern of predictions can result from the design of the training data, the outcome variable selection, or the choice of algorithmic instrument.[172] The individualized hearing model, however, is not well suited to the identification of such systemic problems. Providing an individualized hearing right to all regulated subjects is a good way of providing attention to whether a particular person has been correctly classified. Litigants will not necessarily have incentives, however, to uncover systemic problems (as opposed to highlighting errors in their case). Their retail challenges are hence not necessarily a good way to determine whether there is a problem of inaccuracy-generating flaws in an algorithmic decision-making process.[173] Indeed, the fact that there is error in a specific individual case before an adjudicator is not necessarily evidence of a systemic design problem since most algorithms produce some errors. And once systemic flaws are

169 *Id.* at 127.

170 Coglianese & Lehr, *supra* note 72, at 41.

171 Lehr & Ohm, *supra* note 47, at 656.

172 *See, e.g.*, Altenburger & Ho, *supra* note 73, at 99–100 (exploring how bias in training data can be minimized by the choice of appropriate computational architecture).

173 It is not impossible for individualized hearings to provide a vehicle for reviewing systemic problems. But it hard to see how this could be a cost-effective approach.

rooted out, individualized hearings may be an unnecessarily costly enterprise.

Second, a common assumption of these proposals is that adding human appeals reduces overall rates of false positives and false negatives. But I have argued elsewhere that it is problematic to assume that human decision making is generally more accurate than machine classification or that adding a human appeal to a machine decision will reduce error rates.[174] I was not making a new point. Writing in 1954, the psychologist Paul Meehl compared statistical prediction tools with clinical judgments by trained specialists and came to the conclusion that structured decision making was better (even then) than either humans acting alone or statistical prediction coupled to human review.[175] Recent studies also suggest that adding human oversight to structured (algorithmic decisions) will not always reduce the net volume of false positives and false negatives and instead will often have undesirable, even perverse, effects.[176] While the possibility of a system that successfully integrated post hoc human oversight with machine decisions cannot be ruled out categorically, current proposals that focus on a "hearing officer" are more likely to exacerbate rather than resolve this due process concern.

3. *Testing Algorithmic Design Against Due Process Norms*

In the spirit of Crawford and Schultz, I would instead focus due process analysis on systemic design choices. They, however, provide insufficient detail of how design might compromise due process and how to go about identifying problematic design features. To start filling that gap, I explore here five distinct due process problems that can arise through algorithmic design. All hinge on systemic properties of the machine-learning tool.

---

174 Aziz Z. Huq, *A Right to a Human Decision*, 105 VA. L. REV. (forthcoming 2020) (manuscript at 2), https://ssrn.com/abstract=3382521 [https://perma.cc/G7F8-ZWZC]; *accord* Sharad Goel, Ravi Shroff, Jennifer Skeem & Christopher Slobogin, The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment 3 (Dec. 2018) (unpublished manuscript) (on file with author) (summarizing evidence that additional process can have aggregate negative effects on accuracy).

175 PAUL E. MEEHL, CLINICAL VERSUS STATISTICAL PREDICTION 119–20, 136–38 (1954) (predicting that mechanical predictive methods would outperform clinical ones).

176 Thomas H. Cohen, Bailey Pendergast, & Scott W. VanBenschoten, *Examining Overrides of Risk Classifications for Offenders on Federal Supervision*, 80 FED. PROB. 12, 20–21 (2016); R. Karl Hanson & Kelly E. Morton-Bourgon, *The Characteristics of Persistent Sexual Offenders: A Meta-Analysis of Recidivism Studies*, 73 J. CONSULTING & CLINICAL PSYCHOL. 1154, 1154–56, 1159 (2005).

First, an algorithm might be trained on data that is incomplete, biased, or flawed because of the way that it has been created, selected, or cleaned.[177] Training data produced by state enforcement agencies, such as police or child welfare services, might be shaped by the implicit or explicit bias either of officials or those who provide leads.[178] The result may be an excessive representation of some groups (e.g., racial minorities), not as a consequence of higher misbehavior rates but rather because of the greater propensity of others to report or investigate them. Alternatively, training data might have "black holes" as a consequence of the state's failure to enforce the law in certain locations or against certain populations.[179] Again, the predictable consequences of such flaws is the deviation of predictions from whatever latent construct (e.g., criminality, the risk of benefit ineligibility, or the probability of child abuse) that is the intended object of state intervention. Due process requires at a minimum that an algorithm's designer avoid the unnecessary use of flawed datasets and, where appropriate, take active steps to mitigate training data flaws.[180]

---

177 *See* ALPAYDIN, *supra* note 38, at 40 (describing the use of training and validation data in algorithm design); Michael Mattioli, *Disclosing Big Data*, 99 MINN. L. REV. 535, 561 (2014) (arguing that databases contain errors because of their "sheer size[,] . . . the automatic and indiscriminate information-gathering that is a hallmark of the big data method[, and] . . . errors [that] manifest when error-free data from different sources is merged").

178 A further problem is that "race is such a dominant category in the cognitive field that the 'interim solution' [of using race as a proxy for some other trait of interest] can leave its own indelible mark . . . ." Troy Duster, *Race and Reification in Science*, 307 SCIENCE 1050, 1050 (2005). This means that race might well structure the past deployment of state resources, or patterns of private behavior, in ways that are hard to disentangle from readily available training data.

179 Kate Crawford, *The Anxieties of Big Data*, NEW INQUIRY (May 30, 2014), https://thenewinquiry.com/the-anxieties-of-big-data [https://perma.cc/WV3C-865M].

180 Imagine that an algorithm is accurate for a majority of a regulated population but errs at very high rate for a specific subgroup. Imagine further that this subgroup is not a protected class, defined by race or class. Can members of the nonsuspect class thereby created complain of a due process violation? *Cf.* Ian Ayres, *Outcome Tests of Racial Disparities in Police Practices*, 4 JUST. RES. & POL'Y 131, 139 (2002) (describing this problem). Whether this presents a constitutional problem depends on how costly the subgroup error is to fix for the balance of the population. Where the error cannot be mitigated without introducing greater rates of error elsewhere, for example, due process would not be compromised. But it is worth asking whether it would be minimally rational for the state to continue to use the algorithm in question against the subgroup if it is known that the tool is serially inaccurate. It may be, though, that the state could proffer a reason for not permitting an opt-out (e.g., membership in the subgroup is costly to determine ex ante, and hence it is cheaper to keep the subgroup in). I am grateful to Julian Nyarko for conversations on this point.

Second, an outcome variable may be poorly aligned with the underlying variable of interest, which is commonly termed the "latent construct."[181] For instance, the outcome variable may have been defined in terms of a feature that is not present in the original data. Risk assessment algorithms in the criminal justice space, for example, are designed to predict "dangerousness"—a classification that is not present in the original data.[182] This synthetic classification, however, may not correlate well with the underlying outcome of interest for any number of reasons.[183] The institutional context in which an algorithm is deployed may also influence the fit between an outcome variable and the latent construct. Facial recognition tools are already used to match on police artists' composites.[184] But those composites are likely to be highly imperfect versions of the latent construct of interest: the face of the actual suspect. An algorithm that permits matching on artists' composites therefore introduces a stochastic element associated with a predictably high and racially asymmetrical error rate. Due process might be offended, more generally, by a poor choice of latent construct.

Third, an algorithm's designer might elect a model that is ill-fitted to the policy task at hand. One important election in this regard relates to the important bias/variance trade-off. Model choice, that is, influences a necessary and unavoidable trade-off between bias (how far predictions are from ground truth) and variance (in effect, how much a prediction would vary if the learner was trained on different data sets). There is some evidence that simpler models often perform better than more sophisticated ones because they yield less variance.[185]

---

181 Lehr & Ohm, *supra* note 47, at 679 ("Measurements must be faithful not just to what a variable ostensibly indicates on its face, but also to what underlying construct (also called a latent construct) the data scientist believes it represents.").

182 *Cf.* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 679 (2016) (discussing how algorithms measure creditworthiness, despite there being no direct way to measure creditworthiness).

183 For an argument that "dangerousness" in criminal justice contexts is infected with ideas of race, see Bernard E. Harcourt, *Risk as a Proxy for Race: The Dangers of Risk Assessment*, 27 FED. SENT'G REP. 237, 237 (2015).

184 Scott Klum, Hu Han, Anil K. Jain, & Brendan Klare, *Sketch Based Face Recognition: Forensic v. Composite Sketches*, INT'L CONF. ON BIOMETRICS 3 (2013), https://www.researchgate.net/publication/235701861_Sketch_Based_Face_Recognition_Forensic_vs_Composite_Sketches [https://perma.cc/69JM-UBZW].

185 Pedro Domingos, *A Unified Bias-Variance Decomposition for Zero-One and Squared Loss*, 2000 PROC. 17TH NAT'L CONF. ON ARTIFICIAL INTELLIGENCE 564, 564; *see also* Lehr & Ohm, *supra* note 47, at 697–98 (discussing bias/variance trade-off in algorithmic design).

Depending upon the policy context, different models may be desirable based on how they manage this trade-off. Where precision is less important than consistency as a policy matter, the bias-variance trade-off implies that a model with higher bias might be chosen with the expectation that it will produce a certain rate of errors. Simply examining error rates to condemn or endorse an algorithm without understanding how model choice pertains to policy functions, therefore, may lead a due process analysis astray.

Fourth, an algorithm may be trained on appropriate training data, may initially offer useful predictions on the ground, and then confront cases that defy proper classification. Given the complex and evolving social circumstances in which algorithmic decision tools are likely to work, it is usually needful to evaluate periodically an algorithm's performance to determine that its classifications continue to correspond to the latent variable. This concern may be what Coglianese and Lehr are getting at when they advocate for disclosure of an algorithm's accuracy "in a test data set."[186] But their proposal can be profitably extended to consideration of how an algorithm performs over time on the ground.

Finally, there is a class of cases in which there is no outcome variable available that is well enough correlated to the underlying variable of interest. The algorithm's predictions, therefore, are irrational in the sense of lacking any logical relationship to a legitimate state interest.[187] The problem of irrationality in formal enactments and administrative action has generally been styled as an Equal Protection violation, rather than a Due Process concern.[188] However that problem is

---

186 Coglianese & Lehr, *supra* note 12, at 1187.

187 I should distinguish this point from a similar one made in the literature. Invoking a concern about rationality, for example, Citron has argued that certain decisions "explicitly or implicitly require the exercise of human discretion." Citron, *supra* note 165, at 1302–04. My argument here is different. Citron's argument draws on the well-worn distinction between rules, which are given content before regulated subjects act, and standards, which are given content after regulated subjects act. I think Citron's point is not technically correct as applied to machine learning. There is no technical reason why an algorithmic tool cannot classify new examples and thereby liquidate a standard. The k-nearest neighbor (k-NN) algorithm, for example, classifies new instances by assigning the label that most frequently occurs among the *k* training samples nearest to that query point.

188 Vill. of Willowbrook v. Olech, 528 U.S. 562, 563–65 (2000) (per curiam) (finding that the plaintiffs' allegation that the defendant's actions were "irrational and wholly arbitrary" was "sufficient to state a claim for relief under traditional equal protection analysis," "quite apart from the Village's subjective motivation" (internal quotation marks omitted)); Romer v. Evans, 517 U.S. 620, 631–36 (1996) (holding that Colorado's Amendment 2 "lack[ed] a rational relationship to legitimate state interests").

phrased, it is plausible to say that a constitutional violation is made out when an instrument's outcome variable has no plausible correlation to the underlying outcome of interest.[189]

Teacher evaluations and criminal risk assessments may be cases in point. There is substantial evidence that many available measures of teacher performance, especially student evaluations, are distorted by various improper race- and gender-related biases,[190] or at the very least uncorrelated with measures of learning success.[191] Standardized test data, meanwhile, suffers from vulnerability to gamesmanship by other teachers. As a result, measures of teacher effectiveness based on such scores experience arbitrary fluctuations on a year-to-year basis.[192] Given this, an algorithm trained on either student evaluations or standardized test scores may well be per se invalid on either due process or equal protection grounds (as inaccurate or because it impounds bias).

Or consider the HireVue tool, which may be in use by the Atlanta Public Schools to hire teachers.[193] Apparently, HireVue uses a facial data analytic tool developed by Affectiva, "a leading company in emotion recognition that works in market research and advertising."[194] Even setting aside the doubts that have been raised about the theoretical presuppositions of affect recognition,[195] it is not at all clear how affect, as detected in facial images, is meaningfully predictive of performance as a teacher. Such use of affect recognition in hiring is likely to

---

189 *Cf.* Barocas & Selbst, *supra* note 182, at 715 ("Disputes over the superiority of competing definitions are often insoluble because the target variables are themselves incommensurable.").

190 *See, e.g.*, Friederike Mengel, Jan Sauermann, & Ulf Zölitz, *Gender Bias in Teaching Evaluations*, 17 J. EUR. ECON. ASS'N 535, 563–64 (2019) (finding gender bias in teacher evaluations).

191 *See, e.g.*, Bob Uttl, Carmela A. White, & Daniela Wong Gonzalez, *Meta-Analysis of Faculty's Teaching Effectiveness: Student Evaluation of Teaching Ratings and Student Learning Are Not Related*, 54 STUD. EDUC. EVALUATION 22, 23, 38–40 (2017) (noting that any correlation between student evaluations and learning are flukes instead of due to students' abilities to assess instructors).

192 O'NEILL, *supra* note 157, at 135–40 (critiquing existing models of teacher evaluation).

193 *See supra* notes 110–113 and accompanying text.

194 Patricia Nilsson, *How AI Helps Recruiters Track Jobseekers' Emotions*, FIN. TIMES (Feb. 28, 2018), https://www.ft.com/content/e2e85644-05be-11e8-9650-9c0ad2d7c5b5 [https://perma.cc/X6JR-DG67].

195 *See* Marc A. Cohen, *Against Basic Emotions, and Toward a Comprehensive Theory*, 26 J. MIND & BEHAV. 229, 240 (2005) (arguing that the "research does not support the contention that there is a set of basic emotions"); Michael Price, *Facial Expressions—Including Fear—May Not Be as Universal as We Thought*, SCIENCE (Oct. 18, 2016, 12:33 PM), https://www.sciencemag.org/news/2016/10/facial-expressions-including-fear-may-not-be-universal-we-thought [https://perma.cc/R3RS-LE4Y] (discussing findings that Trobrianders use a gasp to convey anger).

raise a serious question of rationality that, at least in the public sector, has constitutional implications.

Whether criminal risk assessment for bail or probation is ultimately feasible also remains contested. A group of scholars have recently argued that violence risk is so small, even among pretrial detainee populations, that it is statistically infeasible to distinguish the minute number who will go on to commit acts of violence.[196] Moreover, these scholars argue, the training data inevitably used for risk rating is inevitably affected by animus.[197] Other scholars have resisted this conclusion, though,[198] and instruments for predicting violence remain in widespread use.

This list of potential design flaws whereby algorithmic design can go astray is, once again, not intended to be exclusive. Rather, these examples merely illustrate some of the ways in which algorithmic tools can fail to deliver low rates of error.

4. Mathews *and Machine Learning*

The very possibility of specifying ex ante the conditions of due process violation raises an intriguing possibility: Whereas standard applications of the *Mathews* test to agency-based adjudicatory systems can flirt with indeterminacy,[199] its application may be straightforward and predictable in the machine-learning context. Discrete technological design margins can be isolated and then analyzed for their contributions to error rates. Almost fifty years after *Mathews*, that is, technology may be finally making its doctrinal focus empirically tractable. But at the same time, this tractability may also reveal difficulties inherent in the *Mathews* test that until now have been occluded in its judicial application.

To make this more concrete, we can start with the observation that algorithmic tools make different kinds of errors. And it will often be the case that it is technically infeasible to minimize *both* false positives and false negatives.[200] Determining

---

196 Technical Flaws of Pretrial Risk Assessments Tools Raise Grave Concerns 2 (July 17, 2019), https://dam-prod.media.mit.edu/x/2019/07/16/Technical-FlawsOfPretrial_ML%20site.pdf [https://perma.cc/VP6Q-DDF7].

197 *Id.* at 2–4.

198 *See* Goel, Shroff, Skeem & Slobogin, *supra* note 174, at 17 (endorsing risk assessment instruments as superior to clinical predictions).

199 *See* Mashaw, *supra* note 164, at 46 (noting sources of indeterminacy).

200 For papers exploring the kinds of trade-offs implicit in algorithmic design, see Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, & Aziz Huq, *Algorithmic Decision Making and the Cost of Fairness*, PROC. 23RD ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 797, 804–05 (2017); Jon Kleinberg, Sendhil Mullainathan, & Manish Raghavan, *Inherent Trade-Offs in the*

the appropriate mix of false positives and false negatives, then, will require difficult social and normative judgments. These judgments are now often skirted or suppressed in practice. In familiar applications of *Mathews*, these difficult judgments can be elided. In the algorithmic context, however, they become hard to avoid.

As an example of this, consider a binary classification regime, which has false positives and false negatives, rather than a classifier that generates a continuous output variable, which can make errors of degree. The first, binary case is more familiar in a legal context. Algorithmic design recognizes the different value of false positives and false negatives by allowing for different weights to attach to each one.[201] Much as in the civil and criminal trials false positives (negatives) are assigned different implicit weights by varying the burden of proof, that is, so a computational tool can shift the balance between observed false positives and false negatives. But how should due process be defined as between different mixes of false positives and false negatives outside the criminal context? The social value accorded to a false positive as opposed to a false negative in any given situation is a matter of judgment. Wrongful convictions are generally thought to be very costly; erroneous plaintiff verdicts in tort less so.[202] An evaluation of algorithmic due process requires a precise judgment of the relative costs associated with a false negative and a false positive.[203] In respect to bail determinations, employment decisions, and welfare allocations, however, no consensus exists as to the relative values of different error types. In consequence, determining when due process is satisfied will compel an anterior policy debate on the value of different kinds of errors in a given policy domain.

In current practice, a "very common" solution is to assume equal costs from false positives and false negatives.[204] But this seems an implausible global solution. As a result of its displacement, the application of due process will entail difficult judgments about the social costs of various outcomes subject

---

Fair Determination of Risk Scores, 8TH INNOVATIONS THEORETICAL COMPUTER SCI. CONF., at 43:1, 43:4, 43:9, 43:17 (2016), https://arxiv.org/pdf/1609.05807 [https://perma.cc/9L9J-QZLN].

201 Lehr & Ohm, *supra* note 47, at 690–94.

202 *See In re* Winship, 397 U.S. 358, 371–72 (1970) (Harlan, J., concurring).

203 The ordinary application of *Mathews* entails a similar judgment. But algorithmic design allows one to calibrate a performance threshold for accuracy in far more numerically precise terms than litigation would.

204 Hand, *supra* note 57.

to regulation by prediction instruments. And that itself may well be a costly and divisive enterprise.

***

Determining whether a machine-learning tool impinges on due process demands an examination of the fit between quality training data, the learning model, and outcome variable, and the match between the outcome variable and the latent variable. In some ways, this is more difficult than reviewing human decisions; in other ways, it is amenable to more precise analysis.

Provided the fit between training data, learning model, outcome variable, and latent variable is sufficiently tight, a machine-learning tool should pass muster as a matter of due process. I have described those margins, including both choices about training data and methodological choice, in general and nontechnical terms. In many cases, moreover, it will be possible to make judgments about how these design choices were made without access to a classifier's source code.[205] The nature of due process design margins, and their relative availability to ex post scrutiny, has implications for the analysis of remedial frameworks offered in Part III.

B. Equality and Antidiscrimination Norms

The American law of race and gender equality is embodied in the constitutional jurisprudence of the Equal Protection Clause and federal antidiscrimination statutes.[206] Constitutional law, which is my focus here, turns on questions of intent and classification. I explore how these can be adapted to the machine-learning context. I suggest, however, that the equality concerns commonly raised by algorithmic systems in practice are better conceptualized in terms of their impact on pernicious social stratification.[207] In the following, I will focus

---

205 *Cf.* Kroll et al., *supra* note 140, at 638 (discussing the limits of source code review).

206 Note that this standard formulation assumes the identity of equality and antidiscrimination norms. In fact, the conceptual relationship between (different kinds of) equality and antidiscrimination is a complex one. For an excellent treatment, see generally Elisa Holmes, *Anti-Discrimination Rights Without Equality*, 68 MOD. L. REV. 175 (2005) (arguing that anti-discrimination rights do not necessarily require equality).

207 This builds on an earlier critique, but I have tried not to repeat myself here. *Cf.* Huq, *supra* note 10, at 1101–02 (suggesting a need for substantial rethinking of constitutional norms given the diffusion and adoption of machine-learning tools).

on racial equality norms, although many of the points I can make can be transposed to other contexts.

1. *Equal Protection Norms*

The constitutional law of equality takes intent and classification as central analytic terms.[208] Since the mid-1970s, the Supreme Court has defined "the basic equal protection principle" under the Fourteenth Amendment to mean that "the invidious quality of a law claimed to be racially discriminatory must ultimately be traced to a racially discriminatory purpose."[209] It has also held that any occasion upon which "the government distributes burdens or benefits on the basis of individual racial classifications" will lead to the application of strict scrutiny.[210] To survive constitutional scrutiny, a classification's use must be narrowly tailored to serve a compelling state interest.[211] This anticlassification strand of the doctrine is justified on the grounds that racial lines are "divisive" and purportedly rarely relevant to a legitimate state purpose.[212]

The concept of an impermissible "purpose" or intent, however, has not been defined with clarity. It can be construed in

---

208 Statutory antidiscrimination law, in contrast, also includes questions of disparate impact and reasonable accommodation. Noah D. Zatz, *Managing the Macaw: Third-Party Harassers, Accommodation, and the Disaggregation of Discriminatory Intent*, 109 COLUM. L. REV. 1357, 1368–69 (2009). For analyses of how disparate impact liability can be re-articulated for a machine learning context, see, for example, Barocas & Selbst, *supra* note 182, at 701–12 (arguing that the disparate impact doctrine should look for discrimination in data mining); Huq, *supra* note 10, at 1128–33 (arguing for a bifurcated classification rule in algorithmic criminal justice tools).

209 Washington v. Davis, 426 U.S. 229, 240 (1976). Although the intent requirement is now perceived as a conservative formulation, Katie Eyer has persuasively documented how racial progressives advocated for an intent rule through much of the twentieth century as a way to defeat Southern states' efforts to circumvent desegregation rulings. Katie R. Eyer, *Ideological Drift and the Forgotten History of Intent*, 51 HARV. C.R.-C.L L. REV. 1, 4–5 (2016).

210 Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 720 (2007); *see also* Gratz v. Bollinger, 539 U.S. 244, 270 (2003) (describing the use of such classifications as "pernicious" (internal quotation marks omitted)); Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9, 10 (2003) ("[T]he anticlassification . . . . principle holds that the government may not classify people either overtly or surreptitiously on the basis of a forbidden category: for example, their race.").

211 Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 235 (1995) ("Federal racial classifications, like those of a State, must serve a compelling governmental interest, and must be narrowly tailored to further that interest.").

212 Fisher v. Univ. of Tex. at Austin, 136 S. Ct. 2198, 2210 (2016).

several different ways.[213] Consider, for example, a recent racial gerrymandering decision in which the Supreme Court affirmed that "a state law . . . enacted with discriminatory intent" presented a constitutional problem.[214] The Court's reference to "discriminatory intent" might mean several different things. Does it require a showing that legislators responsible for redistricting despised or feared African Americans? What if they simply embraced negative racial stereotypes and hence viewed minorities as less worthy of political influence? Or what if they simply viewed Blacks as "not our people" in a partisan sense? The Court does not say which of these count as "discriminatory intent." Indeed, it is a remarkable feature of Equal Protection jurisprudence that its central term—intent—remains clouded in uncertainty after almost fifty years of service.

Putting this uncertainty to one side, it seems clear that in the modal Equal Protection case, the terms "intent" and "purpose" are typically used to describe the interior psychological disposition or beliefs of a particular individual.[215] To be sure, there are cases in which courts have drawn inferences about the intentions of collective bodies such as legislatures,[216] including racial gerrymandering challenges. But these cases are generally recognized as presenting difficult problems of aggregation and inference because collective bodies do not themselves have intents—only their members do.[217] Even challenges to collective bodies' decisions do not deviate from the baseline psychological model of "intent" as individual belief or disposition insofar as they presuppose the possibility of aggregating individual intents.

2. *Applying Equal Protection Doctrine to Machine Learning*

Difficulties arise in transposing equality doctrine to the machine-learning context. In part, these difficulties track ambiguities in extant applications of that law; in part, they are distinct to this new technology. I consider here how application

---

213 *See generally* Aziz Z. Huq, *What Is Discriminatory Intent?*, 103 CORNELL L. REV. 1211, 1240–63 (2018) (exploring the divergent potential meanings of intent in the constitutional context of antidiscrimination law).

214 Abbott v. Perez, 138 S. Ct. 2305, 2324 (2018).

215 *See, e.g.*, Foster v. Chatman, 136 S. Ct. 1737, 1754 (2016) (invalidating a criminal conviction on Sixth Amendment grounds due to improper preemptory strikes and citing to the prosecutor's "racial animosity" (internal quotation marks omitted)).

216 For a rare example, see Hunter v. Underwood, 471 U.S. 222, 229 (1985).

217 Richard H. Fallon, Jr., *Constitutionally Forbidden Legislative Intent*, 130 HARV. L. REV. 523, 536–37 (2016).

of anticlassification norms and intent-related rules generate difficulties. In this section, I argue that the principal ways in which machine-learning tools raise equality-related concerns are not well captured by anticlassification and intent-focused rules.

Consider first the application of anticlassification rules to the use of race labels in training data. At first blush, the doctrine might be read to suggest that any state use of individuals' race as "an input to [the] system" triggers constitutional concern.[218] The use of race as a "feature" might be seen as analogous to its use as a factor in college applications. In the latter context, the inovocation of race as one factor among many still generates strict judicial scrutiny.[219]

But this line of reasoning may move too fast. For the use of race as a label in machine learning is arguably distinct from its use in college admissions. The latter is public and "divisive"[220] in the way that the technical, often practically indiscernible, use of race in machine-learning systems is not. Moreover, there is a gap between race awareness and impermissible racial classification. Human decision makers employed by the state (such as a police officer or a case worker) are often inevitably aware of race. They are, very simply, immediately presented with phenotypical evidence in the majority of cases. It follows that an official's mere awareness of race raises no constitutional problem. By analogy, it may also be that mere inclusion of race as a feature of training data should not be per se problematic. Rather, such inclusion should be construed to be analogous to the visual accounting for race in quotidian human interactions.[221] Race as a feature is constitutionally problematic only if it influences ultimate decisions in a constitutionally relevant way.

But what counts as a "constitutionally relevant way"? In the intent context, the Court has applied a but-for causation rule.[222] Logically, this should also apply to anticlassification challenges. Applying the but-for causation rule to the ma-

---

218 Barocas & Selbst, *supra* note 182, at 695.

219 Fisher v. Univ. of Tex. at Austin, 136 S. Ct. 2198, 2208 (2016).

220 *Id.* at 2210.

221 For a similar observation in respect to the reliance element of a securities fraud action, see Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J. L. & TECH. 889, 925–26 (2018).

222 *E.g.*, Pers. Adm'r. of Mass. v. Feeney, 442 U.S. 256, 279 (1979) (proof of discriminatory purpose requires showing that government decision maker "selected or reaffirmed a particular course of action at least in part 'because of,' not merely 'in spite of,' its adverse effects upon an identifiable group").

chine-learning context requires courts to determine whether race's inclusion as a feature was a but-for cause of a specific decision. That is, application of a colorblindness rule would lead to a potentially complex technical inquiry into the counterfactual relevance of the race or gender feature.

A race-aware classifier that met this causation requirement, nevertheless, would likely implicate the anticlassification doctrine's concern with "protecting individuals from the harm of categorization by race."[223] As such, it would trigger strict scrutiny. Then, one could ask, how would this standard work, and in particular what would it entail for a racially aware classifier to be narrowly tailored? Because of a statistical phenomenon called "subgroup validity," it is often the case that a failure to include a feature with real-world effects leads to substantial accuracy losses.[224] Excluding race from a learner might have accuracy costs. At some point, the scale of that accuracy loss might be so great that a racial classifier would be (on some view) necessary. It is quite unclear, however, what kind of accuracy loss would be required in order to demonstrate that race's use in a classifier was "narrowly tailored" in a constitutionally adequate way.[225] The Court has never defined clearly what "narrowly tailored" means, nor provided any kind of numerical guidance for its application.[226] This ambiguity already leads to uncertainty in non-algorithmic contexts, where the Court has resolved by failing to provide a precise definition and eliding the definitional question. The leading precedent on point concerns the use of race to propagate diversity in admissions. It rather evasively indicates that narrow tailoring is "simply not susceptible to precise metrics."[227] This solution, though, is not available in the machine-learning context, where an algorithm's designer must assign a numerical value to the accuracy or welfare loss due to making her classifier race- or gender-blind.

---

223 Reva B. Siegel, *From Colorblindness to Antibalkanization: An Emerging Ground of Decision in Race Equality Cases*, 120 YALE L.J. 1278, 1287 (2011).

224 *See* Sam Corbett-Davies & Sharad Goel, The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine-Learning 10 (Aug. 14, 2018), (unpublished manuscript) (providing examples).

225 Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 235 (1995).

226 Richard H. Fallon, Jr., *Strict Judicial Scrutiny*, 54 UCLA L. REV. 1267, 1271 (2007) ("[T]he Supreme Court has never given analytical clarity to the strict scrutiny formula's central concepts of compelling governmental interests and narrow tailoring.").

227 David A. Strauss, Fisher v. University of Texas *and the Conservative Case for Affirmative Action*, 2016 SUP. CT. REV. 1, 16.

Compounding the difficulty in applying the doctrine further, it may well be that the very exercise of using numerical accuracy or welfare-related value as a measure of compliance with the anticlassification norm will strike judges as so inimical to the ethos of constitutional law—so close to a quota—that they would balk at the whole enterprise. In this way, the application of anticlassification rules to machine learning would generate quite novel difficulty.

Application of an intent standard to machine-learning tools can also raise complications.[228] To be sure, it is possible that the designer of a machine-learning tool acts with discriminatory purpose as that term is used in Equal Protection law. But I am unaware of any instance in which animus on the part of an instrument's designers has been credibly alleged.

Discrimination challenges by racial or ethnic minorities based on intent rather than classification, moreover, are notoriously difficult to prove or win.[229] This is so when the official in question openly and repeatedly endorses an illicit motive.[230] Assuming there is no "smoking gun" obtained through discovery or depositions, the task of proving unconstitutional intent will be especially daunting. In particular, when the choice of a certain technical form or a particular set of training data is the basis of the challenge, plaintiffs (especially members of a racial minority) will face an uphill battle.[231] Absent the use of an impermissible classification, plaintiffs alleging intent might argue that a feature was selected because it was "insufficiently rich . . . to assess members of a protected class."[232] Alternatively, they might seek to prove that certain features alone or in juxtaposition have been deliberated selected "as proxies for

---

228 Huq, *supra* note 10, at 1088–94.

229 *See* Russell K. Robinson, *Unequal Protection*, 68 STAN. L. REV. 151, 154 (2016) (contending that "the Supreme Court has steadily diminished the vigor of the Equal Protection Clause in most respects"); Reva B. Siegel, *Foreword: Equality Divided*, 127 HARV. L. REV. 1, 2–3 (2013). Indeed, given how easy it is to discriminate against racial minorities under existing law, there is little or no litigation-related incentive to resort to complex algorithms to cover up impermissible hostility to racial minorities. In contrast, members of racial majorities do not need to demonstrate an illegitimate purpose in affirmative action cases, making the claims more likely to succeed. *See* Siegel, *supra*.

230 *See, e.g.*, Trump v. Hawaii, 138 S. Ct. 2392, 2421–23 (2018) (upholding President Trump's travel ban, despite discriminatory rhetoric, because the ban served legitimate national security purposes); *see also* Aziz Z. Huq, *Article II and Antidiscrimination Norms*, 118 MICH. L. REV. 47, 68–76 (2019) (offering a comprehensive account and critique of that decision).

231 *Cf.* Bathaee, *supra* note 221, at 923–25 (suggesting difficulties in attributing specific outcomes from an algorithm to its designer).

232 Kroll et al., *supra* 140, at 681.

class membership."[233] But I suspect that these arguments will rarely be persuasive in the effort to demonstrate intent.

Further, in many contexts in which the state deploys machine learning, including public benefits and criminal justice domains, race and gender are likely to correlate tightly with other likely features used in training data (such as zip code or socioeconomic outcomes).[234] When there are ready proxies for race or gender effects, a discriminatory state entity can ensure that disfavored groups receive more negative outcomes by including those features in the training data.[235] In criminal justice applications, for example, there are likely to be "plenty of opportunities to associate certain social categories with statistical regularities, stereotypes, and past discrimination."[236] As a result of such collinearity, even a classifier that does not leverage race as a training-data feature is likely to "learn negative associations for certain social labels," including race.[237] That is, discriminatory and nondiscriminatory classifiers may look similar. With the exercise of moderate foresight, therefore, an intentional discriminator can easily skirt liability. Again, this problem is not distinct to the machine-learning context. But the sheer diversity of available features may make it more acute.

In short, the application of anticlassification and intent doctrines (absent a "smoking gun") are likely to generate difficult questions of proof, battles between experts about the purpose of various technical decisions, and few easy resolutions.

3. *Equality and Machine Learning Reconsidered*

Many of the equality-related concerns raised about machine learning, however, do not sound in the register of anticlassification or intent. They instead suggest the need for an alternative normative approach.

A common concern with machine-learning classifiers is their capacity to encode human biases, blind spots, or other-

---

233 *Id.*

234 *See* Barocas & Selbst, *Big Data's Disparate Impact*, *supra* note 182, at 692.

235 A recent paper argues that "proving discrimination will be easier" if algorithms replace human decision makers. Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113, 114 (2019). Where an algorithm designer shapes a model or selects features out of a discriminatory motive, though, this conclusion does not follow.

236 Betsy Anne Williams, Catherine F. Brooks, & Yotam Shmargad, *How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications*, 8 J. INFO. POL'Y 78, 89 (2018).

237 *Id.*

wise normatively troubling assumptions or regularities derived from training data, outcome variables, or other design margins.[238] For example, in 2013, it was shown that a search on Google for typically Black names produced advertisements for arrest records in nearly 90% of cases, while a search for typically white names produced the same sorts of advertisements in less than 25% of cases.[239] In 2019, a different study of a widely used commercial instrument used to recommend care regimes for high-risk patients was flagging equally at-risk Blacks and whites at divergent rates.[240] Black patients, as a result, received fewer interventions despite high morbidity risk. The divergence arose, the study found, because of the instrument's reliance on health care costs as an outcome variable. Recall also some facial recognition tools have errors rates for black women 34.7% higher than those for white men.[241] None of these equality-related concerns are well understood as a worry about either the use of a particular classification or a designer's intent. Perhaps unsurprisingly, the large technical literature on algorithmic bias also eschews a focus on those concepts.[242]

It is possible to generalize from these examples to identify equality-related errors that predictably arise in the machine-learning context but that cannot be easily fit within existing intent-based or anticlassification doctrine. Three examples worth emphasizing are sample bias, feature bias, and label bias.[243]

Sample bias results from nonrandom sampling to create training data. For example, training data for the Allegheny County AFST score arguably reflected bias on the part of mem-

---

238 There are several competing and inconsistent accounts of nondiscrimination in the literature. *See* Huq, *supra* note 10, at 1115–23 (collecting models of fairness).

239 Latanya Sweeney, *Discrimination in Online Ad Delivery*, ACM QUEUE, Mar. 2013, at 1, 12; *see also* SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM 66–80 (2018) (providing examples or race-specific searches that generated derogatory results for Black- but not white-associated terms).

240 Ziad Obermeyer, Brian Powers, Christine Vogeli, & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447, 447 (2019).

241 Buolamwini & Gebru, *supra* note 140, at 11.

242 For a useful recent summary, see Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, & Richmond Y. Wong, *This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology*, 3 PROC. ACM ON HUM.-COMPUTER INTERACTION 119:1, 119:24–26 (2019).

243 These terms are adopted, with some changes, from Corbett-Davies & Goel, *supra* note 224, at 17–19.

bers of the public reporting a risk, with Black families coming under state supervision for more minor infractions than white families.[244] As a result, there were more Black families identified as problematic than white families, leading to distortion in the sample. *Feature bias* occurs if a particular feature assigned to the training data is systematically erroneous because features are mislabeled at different rates across different groups.[245] This might occur in a labor market analysis, for instance, if women are erroneously labeled as less productive as a consequence of biased appraisals.

Finally, *label bias* arises if the designated outcome variable fails to track ground truth equally well for different groups.[246] An outcome variable may evince bias in respect to a specific subgroup where the label is assigned to different social groups at different thresholds. Consider a bail algorithm that is trained using data for which arrest rates are available. If police are more willing to arrest some racial groups rather than others based on the same predicate behavior, then using race as an outcome variable will introduce bias into the data.

None of these problems are well captured by existing Equal Protection doctrine.[247] At a minimum, this suggests that the normative concerns animating the latter are not necessarily identical to the equality-related concerns raised by machine classification. In my view, it is better to recognize that invidious intent and anticlassification do not provide a comprehensive or perspicuous lens to analyze the equality concerns raised by machine-learning tools. While the fashioning of a fully developed alternative to existing equality law is a task that falls beyond my project here, I offer here a very preliminary sketch of

---

244 EUBANKS, *supra* note 75, at 153.

245 *See* Vida Williams, *Combatting Data Bias: Goal, Data, Feature and Model Bias*, MEDIUM (July 23, 2019), https://medium.com/@SingleStoneCX/combatting-data-bias-goal-data-feature-and-model-bias-5aeaf19b83fe [https://perma.cc/PDC6-KAYH].

246 *See* Corbett-Davies & Goel, *supra* note 224, at 17–20.

247 One reason for this is a mismatch with the standard conceptions of discrimination may be a bad fit for the machine learning context. Leading philosophical accounts of discrimination hinge on the notion that certain actions are discriminatory insofar as they manifest disrespect toward a person because they fail to "recognize certain features of . . . persons *qua* persons, such as the intrinsic value of their well-being or the character of their individual autonomy." BENJAMIN EIDELSON, DISCRIMINATION AND DISRESPECT 6 (2015); *see also* DEBORAH HELLMAN, WHEN IS DISCRIMINATION WRONG? 7–8 (2008) (focusing on how "demeaning" action impinges on the "equal worth" of persons). These accounts take as a modal case an interpersonal encounter between individuals in which respect or disdain can be simultaneously manifested and experienced. This is not characteristic of the machine-learning state.

what a reconceptualized approach to equality concerns, at some distance from the current constitutional regime, might look like.

To begin with, it is worth underscoring that the precise nature of "race" remains contested, even among natural and social scientists.[248] Without resolving that disagreement here, it is still possible to observe that race is normatively relevant because it is deployed as a "social fact" by individuals and institutions responsible for critical distributive decisions.[249] As a result of this social usage, race (like gender and disability) has come to be closely correlated with other indicia of disadvantage and exclusion. Thanks to this redundant encoding of race with other measures of exclusion, overt reliance on race or correlated traits (e.g., educational outcomes, residential zip code) often have the effect of strengthening the tendency of resources and opportunities to be distributed in predictably asymmetrical ways. It is the ensuing lopsided diminishment in life chances and material goods for historically marginalized groups that comprises the harm against which equality norms should insulate. A plausible alternative reconceptualization of equality norms for machine-learning instruments therefore focuses on the risk that prediction-driven allocations of benefits or harms amplify the stratifying social effect of race (or, for that matter, kindred classifications such as gender, sexual identity, disability, and ethnicity).

An accounting for such harms in the machine-learning context cannot be done by a mechanical rule against race-consciousness, or by a categorical presumption against prediction. Indeed, it seems to me unlikely, given present levels of racial stratification, that predictive instruments will be able to avoid such harms without some conscientious consideration of the specific mechanisms whereby disadvantage is transmitted over time and space, and (at times) race conscious interventions to disrupt these mechanisms' operation. Where such interventions have social costs (say, by increasing error rates across whole populations), an algorithmic designer must make

---

248 For a documenting of such disagreements, see ANN MORNING, THE NATURE OF RACE: HOW SCIENTISTS THINK AND TEACH ABOUT HUMAN DIFFERENCE 3–8 (2011). For an illuminating debate among philosophers, see generally JOSHUA GLASGOW, SALLY HASLANGER, CHIKE JEFFERS, & QUAYSHAWN SPENCER, WHAT IS RACE?: FOUR PHILOSOPHICAL VIEWS (2019).

249 *See* Eduardo Bonilla-Silva, Reply, *The Essential Social Fact of Race*, 64 AM. SOC. REV. 899, 899 (1999) (internal quotation marks omitted); Mara Loveman, Comment, *Is "Race" Essential?*, 64 AM. SOC. REV. 891, 891 (1999).

decisions about how to trade-off between equity and other goals.

Of course, such trade-offs are politically and normatively controversial. The rise of machine prediction, though, places them in sharp relief. Advances in computational prediction, in other words, are likely to sharpen the conflict between color-blindness and the goal of a social order in which race (or kindred properties) does not define an individual's life course and opportunity set.

C. Privacy

Privacy is a plural not a monolithic concept. It is "complex . . . entangled in competing and contradictory dimensions, [and] engorged with various and distinct meanings."[250] I focus here on one strand: privacy in respect to information, in the sense of an instrumental ability to determine how, and to whom, information held closely by a person is disclosed.[251] In the United States,[252] jurisprudence on informational privacy is far less developed than due process or equality case law. I describe briefly the doctrinal landscape. I then explore the ways in which machine learning can impose distinct harms to informational privacy and ask how a more expansive constitutional or subconstitutional privacy regime might be articulated in response.

1. *Constitutional Privacy Norms*

The Supreme Court has never recognized a free-standing right to informational privacy. In the 1977 case of *Whalen v. Roe*, it assumed arguendo a constitutional entitlement against

---

250 Robert C. Post, *Three Concepts of Privacy*, 89 GEO. L.J. 2087, 2087 (2001). For a useful taxonomy of the various margins of contestation over privacy, see Deirdre K. Mulligan, Colin Koopman, & Nick Doty, *Privacy Is an Essentially Contested Concept: A Multi-Dimensional Analytic for Mapping Privacy*, 374 PHIL. TRANSACTIONS ROYAL SOC'Y A 1, 11 (2016) (distinguishing contests over privacy's foundation, the scope of its protections, the nature of harms involved, and its scope in time and space).

251 Helen Nissenbaum has usefully introduced a distinction between norms of "appropriateness" and "distribution" or "flow" that illuminate "whether [information's] distribution, or *flow*, respects contextual norms of information flow" in a given social sphere. HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE 236 (20010).

252 This is slightly different from the idea of "data privacy" in European law, which is "compromised whenever a data controller processes personal information in a manner that is irrelevant or no longer relevant for the specified purposes for which the information has been acquired." Robert C. Post, *Data Privacy and Dignitary Privacy:* Google Spain, *the Right to Be Forgotten, and the Construction of the Public Sphere*, 67 DUKE L.J. 981, 998 (2018).

the state's improper collection, aggregation, or disclosure of an individual's private information.[253] Although the Supreme Court has never extended *Whalen* to recognize a full-fledged constitutional right to informational privacy, some circuit courts have built on its foundation. A few have suggested that no such right obtains, while others have crafted a cautious doctrinal test for the right.[254] A 2010 precedent appears to read *Whalen* narrowly but conspicuously declined to reject the possibility of a constitutional right to informational privacy.[255] *Carpenter v. United States*, which narrowed the third-party doctrine in the Fourth Amendment context, also recognized a right against government acquisition of private information held by third parties.[256] But third-party doctrine under the Fourth Amendment is analytically distinct from the idea of a free-standing right to control private inferences from data that would otherwise not have been illuminating.

Given the weakness of the constitutional law of information privacy, it is worth looking beyond it to federal and state statutes or regulations. Subconstitutional law, however, is a patchwork. Some federal statutory and regulatory privacy protections generally extend to private actors, but not to federal or state actors.[257] At the subnational level, states such as California, New York, and Massachusetts have imposed data security obligations on large companies, but not state actors.[258] Fur-

253 Whalen v. Roe, 429 U.S. 589, 599 (1977) (describing an "interest in avoiding disclosure of personal matters"); *see also* Nixon v. Adm'r of Gen. Servs., 433 U.S. 425, 457 (1977) (citing the quoted language in *Whalen*).

254 For a discussion of the conflicting lower court precedent on this point, see Lior Jacob Strahilevitz, *Reunifying Privacy Law*, 98 CALIF. L. REV. 2007, 2016–17 (2010).

255 *See* Nat'l Aeronautics & Space Admin. v. Nelson, 562 U.S. 134, 147–48 (2011) ("As was our approach in *Whalen*, we will assume for present purposes that the Government's challenged inquiries implicate a privacy interest of constitutional significance.").

256 138 S. Ct. 2206, 2219–20 (2018) (rejecting application of the third-party doctrine to cell-site locational data). For a useful analysis see generally Alan Z. Rozenshtein, *Fourth Amendment Reasonableness After* Carpenter, 128 YALE L.J. FORUM 943, 947–54 (2019).

257 For instance, acting under the Health Insurance Portability and Accountability Act of 1996, the U.S. Department of Health and Human Services has created by regulation a duty to "[p]rotect against any reasonably anticipated threats or hazards to the security or integrity" of information covered by the statute. 45 C.F.R. § 164.306(a)(2) (2020). Since 1995, the Federal Trade Commission has used its statutory authority to police "deceptive" or "unfair" trade practices to enforce the terms of companies' privacy policies. Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 598–99 (2014) (internal quotation marks omitted).

258 William McGeveran, *The Duty of Data Security*, 103 MINN. L. REV. 1135, 1153–84 (2019).

ther, "a sizeable majority of states have been engaged in privacy enforcement," albeit largely against private actors.[259]

State law, and state officials, may thus fill some of the gaps left by federal law. But it would be wrong to assume that its coverage is comprehensive and systemic, rather than patchy and haphazard. As Lior Strahilevitz has explained in a careful synoptic analysis, this heterogenous approach at both the state and the federal levels means that there may be instances in which a state or federal employee can bring a common-law tort claim of invasion of privacy against an unauthorized governmental disclosure—but whether the employee can will depend on a complex interaction of federal tort liability, immunity doctrines, and state law.[260] Only careful analysis of a particular jurisdiction's applicable federal and state law will reveal whether an action counts as a wrong under either federal or state privacy law.

2. *Privacy Risks from Machine Learning*

The operation of machine learning creates two distinct and new information privacy-related risks. The first involves the power of the state to draw inferences from data that would otherwise not reveal a given private fact. This means "private" information can be acquired without the usual predicate of a constitutionally regulated "search or seizure." Machine learning can implicate different privacy-ousting inferences. One possibility involves "category-jumping" inferences to "reveal attributes or conditions an individual has specifically withheld from others."[261] Examples include the inference of health conditions from spending-related information, or the inference of behaviors or dispositions from health-related data. A second possibility concerns the leveraging of data on one person to draw inferences about an individual who is not present in the dataset. Consider, for example, the genetic databases maintained by both the federal government and all fifty states.[262] Those databases may be searched not only to match those samples, but also to match against "close genetic relatives."[263] Hence, they permit "out of sample" inferences concerning the behavior and location of people who have not come into contact

259 Danielle Keats Citron, *The Privacy Policymaking of State Attorneys General*, 92 NOTRE DAME L. REV. 747, 758 (2016).
260 Strahilevitz, *supra* note 254, at 2017–18.
261 Horvitz & Mulligan, *supra* note141, at 253.
262 Natalie Ram, *DNA by the Entirety*, 115 COLUM. L. REV. 873, 881 (2015).
263 *Id.* at 882–83.

with the criminal justice system. Similarly, consumer genetic platforms, such as GEDmatch and FamilyTreeDNA, contain larger pools of genetic data.[264] Some voluntarily allow law enforcement access. It is likely that the inferential potential of genetic data will increase in the near term. In 2018, researchers used a measure of allele differentiation across the whole genome, called a polygenic risk score, to make impressive population-level predictions of educational and cognitive performance.[265]

A second and distinct form of potential privacy-related harm emerges from a different source. Machine learning depends on the exploitation of large pools of training data. Often held by the state, such pools create a risk of data breaches that impose substantial privacy and pecuniary costs upon individual subjects. In states such as Pennsylvania, officials have even created new data warehouses that collect and house information flows from several, otherwise disparate, state agencies to leverage for predictive ends.[266] Data breaches can result from either negligent or malicious action and come from inside or outside an entity. Studies find a substantial risk of large breaches with the risk rising for any given entity as the amount of data it holds grows.[267]

Breaches of databases can yield not merely unanticipated and socially inappropriate disclosures. As a result of a breach, it is argued, individuals can also suffer "an increased risk of identity theft, fraud, and reputational damage," and immediate "[e]motional distress."[268] It is only because "reliable information regarding the cause, severity and volume of privacy violations is lacking" that there remains uncertainty about both the scale of the problem and the adequacy of legal responses.[269] It

---

264 Natalie Ram, *The U.S. May Soon Have a De Facto National DNA Database*, SLATE (Mar. 19, 2019, 7:30 AM), https://slate.com/technology/2019/03/national-dna-database-law-enforcement-genetic-genealogy.html [https://perma.cc/7L79-3XCG].

265 James J. Lee et al., *Gene Discovery and Polygenic Prediction from a Genome-Wide Association Study of Educational Attainment in 1.1 Million Individuals*, 50 NATURE GENETICS 1112, 1116 (2018) (using polygenic risk scores to explain 11–13% of the variance in educational attainment and 7–10% of the variance in cognitive performance).

266 EUBANKS, *supra* note 75, at 135.

267 Benjamin Edwards, Steven Hofmeyr, & Stephanie Forrest, *Hype and Heavy Tails: A Closer Look at Data Breaches*, 2 J. CYBERSECURITY 3, 4–6 (2016) (reporting findings of an empirical survey of data breaches in the private sector).

268 Daniel J. Solove & Danielle Keats Citron, *Risk and Anxiety: A Theory of Data-Breach Harms*, 96 TEX. L. REV. 737, 745 (2018).

269 Sasha Romanosky & Alessandro Acquisti, *Privacy Costs and Personal Data Protection: Economic and Legal Perspectives*, 24 BERKELEY TECH. L.J. 1061, 1101

seems likely that the diffusion of machine learning across state functions increases the risk of such privacy-related losses above and beyond the risks created by private efforts to collect and analyze individuals' data.

3. *Privacy Rights in the Machine-Learning State*

The range and variation in information privacy harms that can emerge from machine learning obviates the possibility of a single "right to privacy" in that context. Rather than a single right, privacy is better conceptualized as a congeries of entitlements linked by a joint concern with maintaining an appropriate flow of data. Privacy in this context, however, cannot be reduced to a measure of individuated control;[270] the latter is merely one component of a larger repertoire of appropriate responses. I explore three pathways—prohibitions, retail control rights, and privacy "by design"—concluding that the latter is likely most promising despite its shortfalls and limitations.

A first option for responding to machine learning's privacy risks is exemplified by San Francisco's prophylactic bar on facial recognition tools. This is a simple prohibition on the gathering and use of certain kinds of data.[271] I am skeptical, however, that constraints on information acquisition are tenable in the facial-recognition context. The privacy concerns raised by such tools, not least, are unlikely to be addressed successfully by banning public surveillance alone when private surveillance persists. The video surveillance industry throughout the Americas was valued at $3.9 billion in 2016.[272] By the same year, roughly 60% of all cameras sold were network ready. Forty percent of those featured embedded video analytics "as a means to automate the monitoring process and [they] can be particularly effective in proactively identifying events as they happen or extracting information from recorded video."[273]

(2009). For a study of the resulting litigation (which is perforce an unreliable guide to the actual incidence of data breaches), see Sasha Romanosky, David Hoffman & Alessandro Acquisti, *Empirical Analysis of Data Breach Litigation* 11 J. EMPIRICAL LEGAL STUD. 74, 74–75, 93 (2014) (identifying and analyzing more than 230 data breach suits in federal court between 2000 and 2010).

270 *Cf.* Cynthia Dwork & Deirdre K. Mulligan, *It's Not Privacy, and It's Not Fair*, 66 STAN. L. REV. ONLINE 35, 36 (2013) ("[P]rivacy controls and increased transparency fail to address concerns with the classifications and segmentation produced by big data analysis.").

271 *See* Conger, *supra* note 153.

272 IHS MARKIT, VIDEO SURVEILLANCE: HOW TECHNOLOGY AND THE CLOUD IS DISRUPTING THE MARKET 5 (2019), https://cdn.ihs.com/www/pdf/IHS-Markit-Technology-Video-surveillance.pdf [https://perma.cc/4FC7-3PK4].

273 *Id.* at 4, 6.

Even if the state eschews such tools, as in San Francisco, private actors will build databases and pursue recognition-based inferences aggressively. Once private use of these tools is sufficiently pervasive, I am dubious that it will be feasible to maintain a prohibition on state usage of a technology in the face of pervasive private usage. To the public, the latter are likely to seem perverse and otiose—especially in the wake of high-profile crimes or violent crises.

Categorical prohibitions on collection or inference may be more effective, however, in other domains. Since 2008, the Genetic Information Nondiscrimination Act (GINA) has prohibited insurers and employers from relying on genetic data in making coverage or hiring decisions.[274] Because "the paradigmatic GINA claim" arises when an insurer "either drops coverage or hikes up premiums based on a genetic test that reveals a previously unknown health risk," the statute is best understood as a prophylaxis against inferential exploitation of data that, standing on its own, is unilluminating.[275] Bans on certain kinds of machine-learning inference might be justified on privacy grounds, or on the ground that certain kinds of predictions are not properly within the state's authority. GINA, for example, might be justified by the view that biology should not be treated by the state as destiny.[276]

On the other hand, it is hard to see a similar prohibition being extended to state action, since there is some evidence that the creation of DNA databases is associated with meaningful declines in serious crimes, such as murder and rape.[277] Where there are competing social goods that might offset privacy losses, a ban might be implemented with sunset clauses. Temporary measures of this kind would allow regulators to learn how a technology is applied, whether it has greater benefits than costs, and how those costs can be mitigated.

Another alternative is a more narrowly tailored retail right to challenge specific inferences. Use regulation of this sort is already available in the foreign intelligence context,[278] and has

---

274 42 U.S.C. §§ 300gg-53(a)–(c), 2000ff-1(a) (2018).

275 Bradley A. Areheart & Jessica L. Roberts, *GINA, Big Data, and the Future of Employee Privacy*, 128 YALE L.J. 710, 723–24 (2019).

276 *Cf. id.* at 723 (noting concerns about adverse selection in health insurance markets with genetic testing).

277 Jennifer L. Doleac, *The Effects of DNA Databases on Crime*, 9 AM. ECON. J. 165, 166–67, 182–85 (2017).

278 Queries of the bulk metadata collected under Section 215 of the Patriot Act must be supported by "reasonable, articulable suspicion." *In re Application of the FBI for an Order Requiring the Prod. of Tangible Things from [REDACTED]*, No. BR 13–80, at 7 (FISA Ct. Apr. 25, 2013).

been urged by scholars more broadly as a means to regulate government databases.[279]

Yet there is a case for caution before embracing a regulatory reform predicated on dispersed lawsuits by uncoordinated individuals, each challenging a particular use of a machine-learning tool. For one thing, a remedial framework hinging on individualized permissions for machine inferences does not account for the possibility that an official will be able to aggregate insights across several different searches in ways that create new privacy violations. Hence, searching video data for a specific person's movements might constitute a serious privacy invasion only if the officer also has access to that person's internet metadata. A granular system of warrants may thus miss important aggregation-enabled effects.[280]

More generally, in the criminal justice domain, ex ante screens have not proven to be consistently effectual checks on official discretion.[281] The sheer breadth of the modern criminal law lowers the cost of obtaining warrants in the criminal justice context. Similarly, the regulation of machine-learning inferences would be subject to substantive inflation of the justificatory grounds upon which government action is allowed. Given the imperfect performance of the Fourth Amendment's warrant rule in the face of substantive criminal law's inflation,[282] there is no reason for optimism about a parallel ex ante screening rule in the less salient context of machine learning. Instead, the weakness of the present individualized ex ante screening system for criminal searches may be a reason for a more systemic approach in the machine learning context.

A further problem with retail articulation of privacy rights is that individuals seem to be highly imperfect users of protec-

---

279 Emily Berman, *When Database Queries Are Fourth Amendment Searches*, 102 MINN. L. REV. 577, 579–80 (2017) ("[W]hen a database query returns information that the government could otherwise collect only through a Fourth Amendment-regulated means, the Fourth Amendment should regulate that query.").

280 On the other hand, warrants do now impose forward-looking minimization requirements. And in an academic context, institutional review boards can and do place constraints on the combination of empirical data. Enforcing a rule against combinatory actions, however, would require a good deal of tweaking of the present warrant system.

281 Oren Bar-Gill & Barry Friedman, *Taking Warrants Seriously*, 106 NW. U. L. REV. 1609, 1610–11 (2012) ("[W]hat was once a 'warrant requirement' is now a rule so laden with exceptions that it best resembles a piece of Swiss cheese . . . .").

282 Of course, there are conceivable reforms to make warrants more effective. *See, e.g.*, *id.* at 1610–15 (advocating for a clear, revitalized warrant requirement that requires a warrant whenever it is feasible to obtain one). But if those reforms have not taken hold in the ordinary criminal justice domain, should we expect them to take hold in the machine learning domain?

tive tools. One of the distinctive characteristics of privacy harms is the fact that they can arise long after a specific disclosure is made. Retail instantiation of a privacy right assumes that individuals will be able to anticipate and account for temporally distinct harms. It is not clear this is so. Several studies have identified divergent valuations of privacy rights in contractual settings, with variance seemingly motivated by the endowment effects[283] or by an irrational willingness to trade privacy to create a "possibly permanent negative annuity in the future."[284] Cognitive failures of this kind emerge even though the data acquired by platforms and vendors through online transactions has considerable economic value; one estimate suggests that American internet platforms derived $63.8 billion in value from consumers' personal information in 2017 and $76 billion in 2018.[285]

A third possibility beyond bans and retail control rights focuses on building privacy concerns directly into the architecture of a machine learning instrument. There is a range of loosely defined "best practices" for "privacy by design."[286] These require privacy to be "embedded into the design and architecture" of informational systems.[287] Government can implement privacy by design solutions directly or can delegate

283 Alessandro Acquisti, Leslie K. John, & George Loewenstein, *What Is Privacy Worth?*, 42 J. LEGAL STUD. 249, 249–52 (2013).

284 Alessandro Acquisti & Jens Grossklags, *Privacy and Rationality in Individual Decision Making*, 3 ECON. INFO. SECURITY 26, 31 (2005). For similar results, see Kirsten Martin, *Privacy Notices as Tabula Rasa: An Empirical Investigation into How Complying with a Privacy Notice Is Related to Meeting Privacy Expectations Online*, 34 J. PUB. POL'Y & MARKETING 210, 219–21 (2015).

285 ROBERT SHAPIRO & SIDDHARTHA ANEJA, FUTURE MAJORITY, WHO OWNS AMERICANS' PERSONAL INFORMATION AND WHAT IS IT WORTH? 3 (2019), https://www.futuremajority.org/pages/who-owns-americans-personal-information [https://perma.cc/EHA5-B7BX]; *see also* Matthew Crain, *The Limits of Transparency: Data Brokers and Commodification*, 20 NEW MEDIA & SOC'Y 88, 90 (2018) (describing data brokerage as a $200 billion industry). Empirical studies suggest that "[w]hen consumers learn that their data is a tradable asset, they value their data significantly more." Sarah Spiekermann & Jana Korunovska, *Towards a Value Theory for Personal Data*, 32 J. INFO. TECH. 62, 74 (2017).

286 Seda Gürses, Carmela Troncoso, & Claudia Diaz, *Engineering Privacy by Design Reloaded*, 14 CONF. ON COMPUTERS, PRIVACY & DATA PROTECTION, 2011, at 1, 3–4. For the seminal work on this topic, see ANN CAVOUKIAN, PRIVACY BY DESIGN: THE 7 FOUNDATIONAL PRINCIPLES (2009), http://www.privacybydesign.ca/content/uploads/2009/08/7foundationalprinciples.pdf [https://perma.cc/5U2X-B3WH]. The Federal Trade Commission has endorsed privacy-by-design principles. FTC, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES & POLICYMAKERS iii (2012), http://www.ftc.gov/os/2012/03/120326privacyreport.pdf [https://perma.cc/TK9B-X9BL].

287 Cavoukian, *supra* note 286, § 3.

the tasks to private-sector actors who handle sensitive data.[288] Privacy by design operates, as its name suggests, at a system-wide level. One analysis of network security, for example, underscores the need for a "flexible and modular" architecture for holding data.[289] Another catalogs a number of "system[s] . . . designed to detect and prevent the unauthorized access, use, or transmission of confidential information."[290] Data can be classified according to its sensitivity, access can be regulated directly and through encryption, and especially sensitive data can be stored in distributed silos, so no one breach will generate too much damage.[291] Where information is dispersed across numerous physical devices, such as surveillance cameras or the Rapid-DNA "swab in-profile out" box,[292] security against hacks is hard or impossible to achieve through patching, and instead must be integrated in the design and construction stage.[293] The core point is again that privacy, whether a matter of a centralized database or a network of distributed devices, must be hardwired at the design stage. It cannot be effectively supplied at the back end. It is more akin to constitutional structures such as the separation of powers than to a discrete individual right.[294]

To be sure, the strategy of privacy by design is no panacea. In a recent survey, Deirdre Mulligan and Kenneth Bamberger stress the difficulty of "intentionally translating values into design requirements" given cognitive biases and unintended con-

---

288 Kenneth A. Bamberger, *Regulation as Delegation: Private Firms, Decision-making, and Accountability in the Administrative State*, 56 DUKE L.J. 377, 385–86 (2006).

289 Simon Liu & Rick Kuhn, *Data Loss Prevention*, 12 IT PRO. 10, 13 (2010).

290 ASAF SHABTAI, YUVAL ELOVICI & LIOR ROKACH, A SURVEY OF DATA LEAKAGE DETECTION AND PREVENTION SOLUTIONS 10 (2012).

291 Faheem Ullah et al., *Data Exfiltration: A Review of External Attack Vectors and Countermeasures*, 101 J. NETWORK & COMPUTER APPLICATIONS 18, 26–27 (2018); *see also* Lior Arbel, *Data Loss Prevention: The Business Case*, 5 COMPUTER FRAUD & SECURITY 13, 14–15 (2015) (emphasizing the creation of systems for constraining and tracking data access).

292 Rapid DNA analysis is a new technology that allows for DNA testing of buccal swabs to be done at police stations, rather than at a centralized facility. Jacklyn Buscaino et al., *Evaluation of a Rapid DNA Process with the RapidHIT® ID System Using a Specialized Cartridge for Extracted and Quantified Human DNA*, 34 FORENSIC SCI. INT'L 116, 116–17 (2018) (internal quotation marks omitted).

293 Bruce Schneier, *Internet Hacking Is About to Get Much Worse*, N.Y. TIMES (Oct. 11, 2018), https://www.nytimes.com/2018/10/11/opinion/internet-hacking-cybersecurity-iot.html [https://perma.cc/JC8Y-ELKX].

294 *Cf.* Harry Surden, *Structural Rights in Privacy*, 60 SMU L. REV. 1605, 1612–15 (2007) (arguing that if policymakers adhere to the view that privacy rights are coextensive with explicit privacy laws, they may be omitting a significant source of privacy interests).

sequences.[295] This, they argue, is a result of deficiencies in the governmental processes through which privacy by design is realized:

> [E]xisting institutions and processes of democratic and administrative governance have proven to be defective design-war battlefields. They are structurally unsuited to the deliberative decision making [sic] necessary for governance-by-design. No domestic venue exists for the broad conversation about which values to embed in which circumstances. Administrative process frequently fails even to recognize technology design choices as matters of public policy, rather than private choice or government procurement. Agencies generally lack both the technical expertise and the mandate to consider fully the implications of embedding values in design. . . . [A]gency-by-agency decisionmaking [sic] creates downstream ripple effects, prioritizing certain values and precluding reasoned deliberation over others. First movers, particularly those that exercise the greatest sway over the private sector, may co-opt technology to their agencies' particular missions.[296]

In response to these concerns, they offer a series of best practices to mitigate institutional pathologies.[297] Their careful analysis suggests the need for careful institutional design of agencies and departments tasked with the implementation of privacy by design.

In sum, information privacy, like due process and equality, is promoted through the careful design and maintenance of institution-level systems. It is a property of the overall informational architecture in which machine-learning tools are operated, not of any individual act of classification or prediction. No doubt the specific instruments that are best tailored to privacy's production in this context will change as technology shifts and as we move from PC-based applications to phone-based tools to the internet of things (and perhaps thence to mind-AI integration[298]). But it seems probable that the system-level locus of privacy-responsive policymaking will persist.

---

295 Deirdre K. Mulligan & Kenneth A. Bamberger, *Saving Governance-by-Design*, 106 CALIF. L. REV. 697, 710 (2018).

296 *Id.* at 701–02. For criticism of "privacy by design" as ambiguous and an inappropriate delegation of authority to (unrepresentative) engineers, see Ari Ezra Waldman, *Privacy's Law of Design*, 9 U.C. IRVINE L. REV. 1239, 1273 & n.229 (2019).

297 *See generally* Mulligan & Bamberger, *supra* note 295, at 742–80.

298 *Cf.* Alex Knapp, *Elon Musk Sees His Neuralink Merging Your Brain with A.I.*, FORBES (July 17, 2019, 7:41 PM), https://www.forbes.com/sites/alexknapp/2019/07/17/elon-musk-sees-his-neuralink-merging-your-brain-with-ai/

D. Constitutional Norms for Machine Learning: A Summary

My aim in this Part has been to examine how important constitutional values of due process, equality, and privacy are raised by the machine-learning state. Application of those norms implicates not just familiar challenges encountered in the non-algorithmic context but also new problems. In respect to each right, I have suggested a recalibrated account of the relevant norm. In closing, I want to draw attention to a common thread tying these analyses together: When humans interact with algorithmic systems, normative concerns tend to arise because of structural or design decisions that affect many or all users, and not just because of the specifics of particular interactions. *Constitutional norms of procedural due process, equality or privacy, that is, pervasively operate at the system rather than the individual level.* Although this is true in some non-algorithmic contexts, the systematicity of constitutional norms in the machine-learning state creates a strong reason to break from the "liability in tort" model that otherwise dominates adjudication of constitutional rights.

The justifications for adopting a systematic and wholesale, rather than a retail and individualistic, perspective to algorithmic constitutionalism sound in terms of diagnosis, causation, and (relevant to the following Part) remedy. First, from a diagnostic perspective, the identification of individual cases of erroneous decisions provides limited evidence that a particular algorithmic classification system has deviated from due process norms. Nor does the fact that a classification rule tends to rank members of a protected class differently from nonmembers alone bespeak an equality-related problem.[299]

Second, the causes of due process, equality, and privacy violations tend to lie at the level of system design and operation, not the discrete and isolated action of a street-level official. Without taking a systemic perspective that attends to the suite of human design decisions embedded in the algorithm's training data, outcome variable, and method, it will often not be feasible to identify how or why inaccuracies or systemic biases occur. In a like vein, data-breach risk tends to emerge from weaknesses in an information system's architecture. Finally, remedies for due process, equality, and privacy concerns

#1b69a8f74b07 [https://perma.cc/C42P-EX44] (detailing Elon Musk's plan to develop implants to connect human brains with computers).

299 *See* Huq, *supra* note 10, at 1125–32.

are likely incomplete without a systemic perspective. Human appeals from algorithmic decisions may provide due process in the individual case but are likely to increase the overall error rate.[300] Eliminating race from the feature set for an algorithmic tool can lead error rates to spike.[301]

This system-level location of due process, equality, and privacy concerns channels attention to human decisions and elements of algorithmic design remote in time from the immediate contact between a machine and a regulated human subject. As a result, it invites new questions about how, in practice, those norms are to be realized given the dominant "liability in tort" model of constitutional enforcement[302]—questions that are taken up more fully in the next Part.

III
CONSTITUTIONAL REMEDIATION IN THE MACHINE-LEARNING STATE

A well-calibrated remedial architecture for the machine-learning state has two elements. It first requires ex ante rules to force disclosures and generate transparency on the one hand, and to impose accuracy, privacy, and equality-enhancing mandates on the other. Second, it entails the availability of aggregate, rather than individual, litigation remedies after the fact. In other work, I have argued against the idea that a right to a human appeal is an appropriate response to constitutional flaws in a predictive tool.[303] Building on the arguments developed in that earlier article, I posit that aggregate remedies that focus on system-level characteristics of predictive tools provide a more effective means of identifying and correcting design choices that elicit constitutional errors.

The analytic framework employed here draws on a familiar distinction between rules (whose content is established ex ante) and standards (given substance after the fact).[304] This ex

---

300 *See* Huq, *supra* note 174, at 667–71 (developing this argument).

301 *See supra* note 140 and accompanying text.

302 *See supra* note 29 and accompanying text.

303 *See* Huq, *supra* note 174, at 685.

304 Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L.J. 557, 568–77 (1992); *see also* STEVEN SHAVELL, FOUNDATIONS OF ECONOMIC ANALYSIS OF LAW 572–74 (2004) (discussing the fundamental dimensions of legal intervention). A standard is partially specified ex ante, but the full range of relevant considerations, and its precise specification are determined only ex post. For approaches that this parallels, see Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1529, 1552–53 (2019); David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, YALE J. ON REG. 800, 828–36 (2020).

ante/ex post distinction in practice is correlated, somewhat imperfectly, with the choice between regulation by administrative agency and regulation through the common-law system of tort liability.[305] For the sake of simplicity, I assume here that ex ante regulation is done by administrative agencies, while courts undertake ex post review.

Both forms of intervention have familiar strengths. Ex ante regulation trades on the virtues of bureaucratic expertise, predictability, and consistency.[306] Ex post intervention enables private choice by forcing the internalization of potential damage payments and allowing the "parties to calibrate their anticipatory remedial measures."[307] While some scholarship treats these strategies as alternatives, in practice "ex ante and ex post policies are very frequently used jointly."[308] Uncertainty among ex post actors, in particular, can be mitigated by the promulgation of ex ante rules.[309] In the machine-learning context, ex ante regulation can provide off-the-rack templates for disclosure, transparency standards, and design mandates for privacy and equality norms. All these mitigate ex post uncertainty and facilitate diagnosis after the fact. But ex post exposition and review to ensure that constitutional design decisions have been taken and that an instrument has not diminished in accuracy because of brittleness remains a necessary complement.

This convergence is not particularly surprising. It is likely that most policy domains benefit from some mix of ex ante and ex post solutions. The more interesting question is how to calibrate exactly the nature of the instruments used before and after the fact.

305 Richard A. Posner, *Regulation (Agencies) Versus Litigation (Courts): An Analytical Framework*, *in* REGULATION VERSUS LITIGATION: PERSPECTIVES FROM ECONOMICS AND LAW 11, 13–19 (Daniel P. Kessler ed., 2010).

306 Susan Rose-Ackerman, *Regulation and the Law of Torts*, 81 AM. ECON. REV. 54, 54 (1991) (stating that ex ante regulation requires "agency officials to decide individual cases instead of judges and juries; resolves some generic issues in rulemakings not linked to individual cases; uses nonjudicialized procedures to evaluate technocratic information; affects behavior *ex ante* without waiting for harm to occur, and minimizes the inconsistent and unequal coverage arising from individual adjudication").

307 Samuel Issacharoff, *Regulating After the Fact*, 56 DEPAUL L. REV. 375, 380 (2007).

308 Charles D. Kolstad, Thomas S. Ulen, & Gary V. Johnson, Ex Post *Liability for Harm vs.* Ex Ante *Safety Regulation: Substitutes or Complements?*, 80 AM. ECON. REV. 888, 888 (1990) (emphasis omitted).

309 *Id.* at 889; *see also* Steven Shavell, *A Model of the Optimal Use of Liability and Safety Regulation*, 15 RAND J. ECON. 271, 271 (1984) ("[I]t is often socially advantageous for the two means of controlling risk to be jointly employed—for parties to be required to satisfy a regulatory standard and also to face possible liability.").

Even assuming this need for ex post enforcement through litigation, questions remain about the form of litigated oversight. I emphasize here the virtues of aggregate litigation over retail challenges to the outcomes of specific cases. Aggregate challenges (such as class actions, facial challenges, and the like) usefully direct attention to system-wide causes of constitutional harm. They invite remedies fashioned to account for the interests of all regulated subjects—and not, say, instruments that improve on accuracy for a subset of the regulated population while increasing errors for a majority. This aggregate/retail distinction is not the sole important question of remedial decision choice (and is surely not important *only* in this context). But I focus on it because of its singular importance in the machine-learning context.

A. Regulating Algorithms

Administrative agencies have long been "key actors responsible for implementing congressional commands contained in statutes."[310] In comparison to legislators and courts, agencies boast comparative institutional advantages in expertise and responsiveness.[311] Ex ante regulation is possible by both federal and subnational agencies. States such as California are enacting statutory protections of privacy that will impinge on the way in which private actors can deploy machine learning.[312] Municipalities such as Seattle and Santa Clara have

---

310 Bertrall L. Ross II, *Embracing Administrative Constitutionalism*, 95 B.U. L. REV. 519, 527 (2015); *see also* Sophia Z. Lee, *Race, Sex, and Rulemaking: Administrative Constitutionalism and the Workplace, 1960 to the Present*, 96 VA. L. REV. 799, 801 (2010) (defining administrative constitutionalism as "regulatory agencies' interpretation and implementation of constitutional law"); Gillian E. Metzger, *Administrative Constitutionalism*, 91 TEX. L. REV. 1897, 1900 (2013) (describing administrative constitutionalism as "encompass[ing] the elaboration of new constitutional understandings by administrative actors");. Of course, this might change if constitutional doctrine changes. *Cf.* Gundy v. United States, 139 S. Ct. 2116, 2131–32 (2019) (Gorsuch, J., dissenting) (casting doubt on rule-making delegations to federal agencies).

311 Margaret H. Lemos, *Special Incentives to Sue*, 95 MINN. L. REV. 782, 786–87 (2011).

312 *See, e.g.*, Assemb. B. 375 (Ca. 2018) ("[G]rant[ing] . . . consumer[s] a right to request a business to disclose the categories and specific pieces of personal information that it collects about the consumer, the categories of sources from which that information is collected, the business purposes for collecting or selling the information, and the categories of 3rd parties with which the information is shared."); Dipayan Ghosh, *What You Need to Know About California's New Data Privacy Law*, HARV. BUS. REV. (July 11, 2018), https://hbr.org/2018/07/what-you-need-to-know-about-californias-new-data-privacy-law [https://perma.cc/49DF-ADV4] (summarizing the background and effects of California's Consumer Privacy Act).

enacted regulations covering not only the collection but also analysis of surveillance data.[313] These examples are unlikely to prove isolated. To the contrary, interjurisdictional diffusion, imitation, and competition likely will generate healthy rates of regulatory innovation even absent federal action.

Ex ante regulation can be used to create substantive standards or to create a disclosure regime. I address each of these possibilities in turn.

1. *Substantive Regulatory Interventions*

The most common ex ante regulatory intervention relevant to machine learning in nonpublic hands is privacy by design. Both the European Union and the federal government have adopted mandates of that kind.[314] Scholars have devoted considerable attention to refining privacy-by-design principles.[315] I will focus here on regulating for equality. This is a useful focus because the regulatory focus on privacy to date has made equality values more costly to enforce because it has deprived regulators and private parties of information necessary to identify discriminatory phenomena.[316] For example, a 2019 Illinois statute regulating the use of machine learning in hiring decisions mandates the destruction of video data within thirty days of an interview upon an interviewee's request—a measure that likely makes it more difficult to ascertain ex post whether unlawful discrimination may have occurred in the hiring process.[317] As legislators and agencies consider how public uses of machine learning are managed, greater attention to computational infrastructure conducive to equality norms is thus

---

313 Seattle Mun. Code § 14.18.010 (Wash. 2017) (regulating "any electronic data collected, captured, recorded, retained, processed, intercepted, or analyzed by surveillance technology acquired by the City or operated at the direction of the City"). Similar measures include Santa Clara County, Code of Ordinances § A40-7(c) (Cal. 2020).

314 *See* FTC, *supra* note 286, at 22–34; Commission Regulation 2016/679, 2016 O.J. (L 119) 1 (EU).

315 *See, e.g.*, WOODROW HARTZOG, PRIVACY'S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW TECHNOLOGIES 12 (2018) (offering a framework for law and policy that uses privacy by design to regulate consumer protection and surveillance); Mulligan & Bamberger, *supra* note 295, at 740–80 (proposing a new institutional, technological, and conceptual framework to preserve privacy-by-design); Waldman, *supra* note 296, at 1266–85 (using products liability to answer privacy-by-design's open questions).

316 Mulligan & Bamberger, *supra* note 295, at 728 ("Limiting the availability of attributes like race, gender, and nationality can limit blatantly intentional discrimination but confounds efforts such as this to root out more invidious forms of discriminatory profiling."); *accord* Dwork & Mulligan, *supra* note 270, at 37.

317 Artificial Intelligence Video Interview Act, 820 ILL. COMP. STAT. 42/1 (2020).

useful. In that spirit, this section outlines an equality-related regulatory intervention—a mandate to adopt the "best feasible" nondiscriminatory algorithm. This idea, I should note in advance, need not be limited to equality norms, but might also have due process and privacy applications.

One regulatory mandate worth exploring works by analogy to the "Best Available Technology" (BAT) rules employed in several federal environmental statutes.[318] The gist of the idea is that regulating agencies would mandate a BAT requirement for nondiscriminatory (fair) algorithms (although it is possible to engage the same mandate in respect to security against data breaches).

Under the Clean Water Act, for example the EPA determines the "best practicable control technology" by accounting, inter alia, for "the total cost of application of technology in relation to the effluent reduction benefits to be achieved from such application" and "the age of equipment and facilities involved, the process employed, the engineering aspects of the application of various types of control techniques, process changes, . . . [and] environmental impact."[319] BAT mandates of this ilk allow the agency to derive an appropriate regulatory standard from the observed distribution of industry practices.[320] Closer to the context at hand, they have been proposed as a liability rule for websites' responsibilities respecting copyright enforcement.[321]

BAT rules might be implemented in a number of different ways. For example, they might be framed in general terms so as to impose a burden on regulated actors to select or develop instruments that minimize a set of race- or gender-related costs and benefits or to maximize certain outcomes. Rather than directing those actors to employ a preselected instrument, the mandate would leave it to courts to ascertain what counted as a BAT through after-the-fact litigation. This approach leverages the possibility that regulated actors are better positioned than agencies to identify and develop mechanisms for optimizing over costs and benefits.

---

318 *See, e.g.*, 33 U.S.C. § 1311(b)(2) (2018) (requiring Best Available Technology economically achievable for toxic pollutants under the Clean Water Act).

319 *Id.* at § 1314(b)(1)(B).

320 Jonathan S. Masur & Eric A. Posner, *Norming in Administrative Law*, 68 DUKE L.J. 1383, 1396–98 (2019); *see also* Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972, 2024–25 (2017) (offering this suggestion in respect to machine-based testimony).

321 Lital Helman & Gideon Parchomovsky, *The Best Available Technology Standard*, 111 COLUM. L. REV. 1194, 1217–18 (2011).

Alternatively, an agency might simply promulgate an open-ended "list of best available technologies . . . ex ante" from which regulated entities would select.[322] This pathway would place a burden on the regulating agency to identify equality-favoring innovations ex ante. The agency might derive this information from observation of private market behavior, or alternatively, through an information-revelation mechanism such as a system of prizes or research grants.[323] Finally, a BAT for constraining discriminatory effects might entail the crafting of an equality term that can be included in a classifier equation.[324] Of course, any of these regulatory approaches requires the agency to define ex ante the form of (racial or gender) equality it deems important even if the burden of technical design that falls on the agency would otherwise vary.

BAT mandates of this form, in sum, illustrate the kinds of substantive mandates that can be used to elicit ex ante salutary forms of algorithmic action. The example, though, is not meant to be exhaustive. To the contrary, I offer it to suggest the potential of regulatory mandates, with the expectations that others can and should be imagined.[325]

2. *Transparency and Disclosure Mandates*

Another pathway for ex ante regulation focuses on disclosure of various sorts—or forms—of what has come to be known as transparency and explainability in algorithmic design. I be-

---

322 *Id.* at 1224.

323 For the relative merits of prize mechanisms, see Brian D. Wright, *The Economics of Invention Incentives: Patents, Prizes, and Research Contracts*, 73 AM. ECON. REV. 691, 696–700 (1983).

324 This is suggested in an unpublished paper. *See generally* Michele Samorani, Shannon L. Harris, Linda Goler Blount, Haibing Lu, & Michael A. Santoro, Overbooked and Overlooked: Machine Learning and Racial Bias in Medical Appointment Scheduling 15–16 (Oct. 9, 2019) (unpublished manuscript), https://ssrn.com/abstract=3467047 [https://perma.cc/E8E5-YRJP]. The proposal, however, is novel and should be regarded as only a possibility absent further scrutiny.

325 In the privacy context, for example, one mandate might focus on minimizing the risk of deanonymization. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1716 (2010). A 2011 comprehensive metastudy of health-related data acknowledged reidentification risk and concluded that it was "insufficient" to draw strong conclusions about the magnitude of such risk. *See generally* Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, & Bradley Malin, *A Systematic Review of Re-identification Attacks on Health Data*, PLOS ONE, Dec. 2011, at 1. However, more recent work underscores the possibility of embedding privacy-protective design features into data to prevent reidentification, including the exclusion of certain features and perturbation of the data. *See generally* Khaled El Emam, Sam Rodgers, & Bradley Malin, *Anonymising and Sharing Individual Patient Data*, 350 BMJ 1 (2015).

gin by offering a cautious note about the ambiguous meaning and potential costs of transparency. I then explore specific ways in which these difficulties can be resolved. Finally, I identify some specific disclosure mandates that facilitate important ex post judgments about constitutional norms, even though these are not well described as "transparency" mandates.

Despite a recent "resurgence" of interest in "explainable artificial intelligence," the precise meanings of that term and its cognate "transparency" remain hotly contested.[326] The former term has even been criticized as a "suitcase word[ ]" that "pack[s] together a variety of meanings" but that "holds no universally agreed-upon meaning."[327] A threshold, and critical, ambiguity concerns the threshold object of the exercise. A disclosure mandate might focus either on "the mechanism by which the model works" or, alternatively, on a justification or an explanation of a specific classification decision.[328] This is the difference between a request for a global explanation (i.e., providing a covering law that characterizes the algorithm's work) and a local explanation (focused on a specific instance).[329]

Popular writing often seems to assume that machine learning is unavoidably inscrutable.[330] And indeed, it is the case that many forms of machine-learning architectures are so complicated that their manner of computing outcomes, or their

---

326 Tim Miller, *Explanation in Artificial Intelligence: Insights from the Social Sciences*, 267 ARTIFICIAL INTELLIGENCE 1, 1–2 (2019). For another survey that underscores the breadth of the term, see Michael Gleicher, *A Framework for Considering Comprehensibility in Modeling*, 4 BIG DATA 75, 77–84 (2016).

327 Zachary C. Lipton & Jacob Steinhardt, *Troubling Trends in Machine-Learning Scholarship*, 17 ACM QUEUE, Jan.–Feb. 2019, at 1, 15.

328 Zachary C. Lipton, *The Mythos of Model Interpretability*, ACM QUEUE, May–June 2018, at 1, 12–13; *cf.* Coglianese & Lehr, *supra* note 72, at 20–22 (distinguishing "fishbowl" transparency, which is transparency into what the government has done, from "reasoned" transparency, which focuses on the reasons for action). Selbst and Barocas distinguish between inscrutability (pertaining to how something works) and nonintuitiveness (why it works that way). Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1089–91 (2018). These margins both concern the choice of method and not the result. Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1089–91 (2018). These margins both concern the choice of method, and not the result.

329 Amina Adadi & Mohammed Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, 6 IEEE ACCESS 52138, 52147–48 (2018) (drawing the global/local distinction).

330 *See, e.g.*, Knight, *supra* note 15 ("We've never before built machines that operate in ways their creators don't understand. How well can we expect to communicate—and get along with—intelligent machines that could be unpredictable and inscrutable?"). Knight, to be sure, recognizes that he is discussing only a subset of machine learning.

design, cannot be easier conveyed in a nontechnical form. This is acutely so for deep-learning instruments.[331] In 2015, for example, Microsoft developed a prize-winning convolutional neural network called ResNet.[332] Not only did ResNet have 152 layers of neurons in its network, it also used a device called skip-connections, which allow neurons in an "outer" layer to feed directly into neuron layers much deeper in the network's architecture. Accounts of ResNet suggest that there is no easy way to "explain" how the network operates to a nonspecialist, or to retrace the computational steps needful to reach a particular outcome. If "transparency" is understood to demand an account of how ResNet works in its particular that is legible to a lay person, it may well be a fool's errand.

But ResNet is not necessarily typical of the models currently in common state use. The assumption that all machine-learning models are as impenetrable as ResNet is also flawed. For there are other methods, such as decision trees and linear models, that are far more "easily understandable and interpretable for humans."[333] At the global level, therefore, the available scope for explanation is a function of the choice of algorithmic method. The most sophisticated (and hence effective) algorithms in usage now, deep learning instruments, tend to be the most difficult to represent because of their scale, their use of distributed representations, and the iterative nature of their computations.[334] While there is research ongoing on rendering deep learning instruments more intuitive through a combination of expository tools,[335] global-level transparency mandates focused on how a specific method operates are likely to require a trade-off between competing normative ends of transparency and accuracy. At times this trade-off can be avoided. One way to mitigate it, for example, is to seek "simple rules" that per-

---

331 Marcus, *supra* note 60, at 10–11.

332 KELLEHER, *supra* note 43, at 170.

333 Riccardo Guidotti et al., *A Survey of Methods for Explaining Black Box Models*, 51 ACM COMPUTING SURVEYS 93:1, 93:7 (2018).

334 KELLEHER, *supra* note 43, at 243–44; Adadi & Berrada, *supra* note 329, at 52145.

335 Chris Olah and colleagues, for example, have suggested that "disparate techniques now come together in a unified grammar, fulfilling complementary roles in the resulting interfaces . . . [that] allow[ ] us to systematically explore the space of interpretability interfaces, enabling us to evaluate whether they meet particular goals." Chris Olah et al., *The Building Blocks of Interpretability*, DISTILL (Mar. 6, 2018), https://distill.pub/2018/building-blocks/ [https://perma.cc/T2KX-SRQL]. They use this composite method to offer explanations of deep learning tools.

form (almost) as well as complex instruments yet are more readily comprehensible.[336]

Within these constraints, an explanation of a classification outcome—why was this person jailed, or that benefit denied?—might proceed in a number of different ways. Like global explanations, outcome-specific explanations can be more or less feasible depending on how they are conceptualized. An outcome could be explained in terms of its designer's goals: *x* result was reached because the algorithm was designed to do *p*. It could alternatively index the specifics of an instrument's technical architecture (say, the manner in which hyperparameters were calibrated).[337] A third form of explanation focuses on causality. To "explain" a specific outcome might thus be to offer a causal explanation—a formulation that might elide with a method-focused definition of transparency, or that might run into difficulty because of the noncausal quality of much machine-learning inference.

In contrast to these approaches—each of which raises technical or conceptual difficulties—recent studies of explanation in the machine-learning context instead suggest that the most commonly observed demand from human users is one for "contrastive" explanations. These "do not explain the causes for an event per se, but explain the cause of an event relative to some other event that did not occur.[338] That is, they give an answer to the question "why *x* and not *y*." A demand for a contrastive explanation entails the identification of counterfactuals in which a minimal number of features are changed to reach a different classification, or a justification that links that outcome to some underlying policy judgment or latent variable.[339] Transparency of this kind is a tractable design option in many cases. But which of these implementation mechanisms is appropriate will depend on the specific normative questions raised by algorithmic decision making in a given context.[340]

---

336 *See, e.g.*, Jung et al., *supra* note 41 (exploring the availability of "fast, frugal, and clear" decision procedures across a range of domains).

337 That is, terms set by human judgment rather than being computed by the machine itself.

338 Miller, *supra* note 326, at 9.

339 CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING: A GUIDE FOR MAKING BLACK BOX MODELS EXPLAINABLE 37–38, 241–43 (2020).

340 Menaka Narayanan et al., How Do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation 1–3, 15 (Feb. 5, 2018) (unpublished manuscript) (discussing why different kinds of explanation differ, and how to craft effective responses).

In addition to these decision-specific options, there is a range of more specific disclosure mandates to facilitate ex post accounting. I offer three examples of these.

First, an algorithmic decision should be accompanied by a "datasheet" that records the choices and manipulations of training data, and the "composition, collection process, recommended uses, and so on" of the raiding data.[341]

Second, an algorithm should be designed for "auditability . . . to enable third parties to probe and review the behavior of an algorithm."[342] At a most basic level, this might be done through inclusion of an application programming interface (API) that facilitates downstream review even without access to the underlying algorithm.[343]

Finally, cryptographic commitments embedded in an algorithm's code are a way of ensuring that the same, known decision rules are applied to all regulated subjects.[344] A related possibility, developed by the Open Algorithms project of Imperial College London and the MIT Media Lab, is the use of blockchain as a record to log the manner in which an algorithm is used across particular cases.[345] A similar possible design mandate with the ambition of enabling proof ex post would require an algorithm to produce "a tamper-evident record that provides non-repudiable evidence of all nodes' actions."[346]

---

341 Timnit Gebru et al., Datasheets for Datasets, 2 (Jan. 15, 2020) (unpublished manuscript).

342 Nicholas Diakopoulos & Sorelle Friedler, *How to Hold Algorithms Accountable*, MIT TECH. REV. (Nov. 17, 2016), https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/ [https://perma.cc/3VFU-ZTJM].

343 It is possible to access a black-boxed algorithm via an API to test how certain features (e.g., protected-class membership) influences outcomes without disclosing the algorithm's operating rules. Philip Adler et al., *Auditing Black-Box Models for Indirect Influence*, 54 KNOWLEDGE & INFO. SYSTEMS 95, 96–97 (2018).

344 Kroll et al., *supra* note 140, at 665–67 (describing a "cryptographic commitment," a digitally generated, tamper-proof certification, that assures that "(1) [a] particular decision policy was used and (2) . . . particular data were used as input to the decision policy"). Another precommitment device is the zero-knowledge proof, which can be used to prove that a certain decision policy was actually used without revealing its contents. *Id.* at 668.

345 Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, & Patrick Vinck, *Fair, Transparent, and Accountable Algorithmic Decision-making Processes*, 31 PHIL. & TECH. 611, 622–24 (2018) (describing the implementation of the Open Algorithms project).

346 Andreas Haeberlen, Petr Kouznetsov, & Peter Druschel, *PeerReview: Practical Accountability for Distributed Systems*, 2007 ACM SIGOPS OPERATING SYSTEMS REV. 175, 175; *accord* Deven R. Desai & Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 10–11 (2017).

None of these options ought to be impeded by trade secrecy claims on behalf of algorithms' creators.[347] A regulatory agency should mandate that certain parameters and hyperparameters be disclosed alongside a machine's operation. For due process purposes, this might include the nature and origins of the training data, any constraints imposed upon rules that could be learned from that data, the outcome variable, and the latent construct. It is difficult to see how any of these disclosure obligations would impinge upon intellectual property interests in algorithmic design, even on the assumption that such an interest was a substantial one, given the availability of a protective order. Even where a vendor who has sold the state an algorithmic system does claim intellectual property protection, a regulator could reasonably compel the vendor to make public sufficient detail to understand how historical data is translated into prediction or prescription. Agencies not only have clear power to condition access to state contracts on such disclosure, they can appeal to the publicity-oriented justification of intellectual property law itself.[348]

***

Because the decisions relevant to those norms are often embedded in the threshold development and design of a machine-learning system, regulators are well positioned to generate mandates and constraints that conduce to constitutional compliance. Indeed, a takeaway from my analysis is that there is a wide array of ex ante tools available to regulators wishing to promote constitutional norms in the machine-learning state. The taxonomy offered here is not an exhaustive guide to how such regulation should be framed. It rather presents a first step in developing needful regulatory frameworks for promoting a machine-learning state under the rule of law.

B. Litigating the Constitutionality of Algorithms

Ex ante regulation is necessary, but is not sufficient, to promote constitutionalism in the machine-learning state. Designers of a machine-learning system cannot be certain before the fact of how their instrument will perform across all conceivable circumstances. Learned rules can and do prove brittle in

347 *See* Wexler, *supra* note 145 (describing the problem with creators protecting their algorithms with trade secrecy claims).

348 Mark A. Lemley, *The Surprising Virtues of Treating Trade Secrets As IP Rights*, 61 STAN. L. REV. 311, 332–33 (2008).

the teeth of unexpected phenomenon.[349] Designers of a machine-learning system, even if subject to robust ex ante regulation, may also fail to install or maintain appropriate protections for constitutional norms. Privacy-protective software patches, for example, might not be timely installed. Hardware obsolescence may not be mitigated. A loose fit between the outcome variable and the latent construct of interest may slip into the design. As a result, some form of ex post litigation is necessary even with ex ante regulation in place.

The optimal litigation form for enforcing constitutional norms in the machine-learning state is wholesale and not retail. It takes the algorithmic system's operation as the relevant transactional frame. It offers injunctive relief aimed at correction and improvement of that system's operation as a remedy. It should not aim to generate damages or even categorical negative injunctions that prohibit machine learning in all circumstances, or even opt-outs for specific, select plaintiffs without any regard to how the majority of regulated subjects are treated.[350] Litigation's ambition, therefore, should be understood in terms of systemic amelioration in line with the wholesale nature of due process, equality, and privacy norms.

A suit to enforce constitutional norms against an algorithmic governance tool will perforce focus on the tool's system-level operation. Due process challenges under *Mathews* will usually turn on one of the ways (discussed above) in which algorithmic architecture can generate substantial numbers of false positives or false negatives.[351] Equality challenges hinging on either intent or classification will centrally concern the choices of training data, features, and outcome variable (although the way in which those parameters are analyzed remains up in the air).[352] And privacy litigation will tend to focus on system-level vulnerabilities of software or hardware, and failures to implement privacy by design.[353] Regulatory mandates along certain design margins, such as transparency requirements, cryptographic commitments, and zero-day proofs can facilitate litigation by rendering predictable litigants' access to important empirical and technical details. And a burden-shifting mechanism, akin to that used in disparate impact

349 *See supra* notes 99–102 and accompanying text.

350 *See* Huq, *supra* note 174, at 628–29.

351 *See supra* section II.A.2.

352 *See supra* section II.B.2.

353 *See supra* subpart II.C.

litigation,[354] can be used to weed out insufficiently robust design choices along all three margins.

Constitutional litigation in this vein can be filed either by private or public plaintiffs. A public agency would file suit against a coordinate body within government. Such suits can be observed at both the federal[355] and the state level.[356] States also have "parens patriae" standing to vindicate "quasi-sovereign" interests, which is understood to include a "general interest" in the welfare of its citizens of the sort that a state might try "to address through its sovereign lawmaking powers."[357] The latter might be relevant when constitutional interests are vindicated best through a suit against a private party acting in coordination with the state. A parens patriae suit might be brought, for example, against the supplier of algorithmic software or the hardware on the ground that it (say) created an improper risk to state residents' privacy interests.

Such suits have not to date been brought. Even if they emerge, it seems likely that public enforcement of constitutional norms in the machine-learning context will remain at undesirably low levels. Agencies operating under a state or federal aegis have strong incentives to settle their disputes internally rather than in the court. At present, the necessary institutional infrastructure for the robust enforcement of due process, equality and privacy norms detailed in Part II simply does not exist. In its absence, it seems likely that private litigation will continue to play an important role in trying to vindicate constitutional norms in the machine-learning state.[358]

The obvious form that private enforcement could take is the class action suit in state or federal court. The Supreme Court has recently restricted state courts' jurisdiction to adjudicate national class actions.[359] But state courts remain able

---

354 *See* 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2018) (setting forth burden shifting test for Title VII).

355 Daniel A. Farber & Anne Joseph O'Connell, *Agencies as Adversaries*, 105 CALIF. L. REV. 1375, 1415 (2017) (documenting cases).

356 *See, e.g.*, Va. Office for Prot. & Advocacy v. Stewart, 563 U.S. 247, 261 (2011) (permitting *Ex parte Young* action by an independent state agency against a coordinate agency).

357 Alfred L. Snapp & Son, Inc. v. Puerto Rico *ex rel.* Barez, 458 U.S. 592, 607 & n.14 (1982) (emphasis omitted).

358 For an analogous argument in the antitrust context, see HERBERT HOVENKAMP, THE ANTITRUST ENTERPRISE: PRINCIPLE AND EXECUTION 58–63 (2005).

359 *See* Bristol-Myers Squibb Co. v. Superior Court, 137 S. Ct. 1773, 1783–84 (2017). For a useful discussion of the case's effects, see generally Andrew D. Bradt & D. Theodore Rave, *Aggregation on Defendants' Terms:* Bristol-Myers Squibb *and the Federalization of Mass-Tort Litigation*, 59 B.C. L. REV. 1251, 1281–1306 (2018).

to resolve challenges to state-level policies implemented by state officials. Such suits have been lodged, for example, to challenge deficiencies in the funding of public defense offices and other criminal justice dysfunctionalities.[360] And as noted, there is already a scattering of suits challenging the use of machine learning and similar tools in public benefits, teacher evaluation, and bail contexts.[361] A thousand more flowers, so to speak, should bloom.

Suits challenging algorithmic governance have yielded a range of reforms. In Houston, the challenge to the EVAAS teacher evaluation system led to the school district abandoning algorithmic assessment.[362] In the challenge to the Arkansas benefits system described earlier, litigation revealed that "a third-party software vendor implementing the system[ ] [had] mistakenly used a version of the algorithm that didn't account for diabetes issues," and forced the state to correct the flaw.[363] And in an Idaho suit challenging a benefits algorithm, plaintiffs "work[ed] with the Idaho Department of Health and Welfare to develop a new model."[364] The settlement ultimately accepted by the Idaho district court contained a twenty-four-step process for evaluating and recalibrating the benefits process.[365] None of the cases I have identified ultimately led to a damages award. This militates against the concern that legal challenge will generate disabling liabilities for state and municipal actors out of proportion to their fault.[366] These examples suggest that

---

360 *See, e.g.*, Hurrell-Harring v. State, 930 N.E.2d 217, 219–20 (N.Y. 2010) (challenging that the state's underfunded public defenders deprive indigent defendants the right to Assistance of Counsel); Kuren v. Luzerne Cty., 146 A.3d 715, 718 (Pa. 2016) (same); *see also* Pub. Def., Eleventh Judicial Circuit v. State, 115 So. 3d 261, 265–66 (Fla. 2013) (stating that public defenders successfully moved to withdraw from nonfelony cases, citing a lack of resources); Phan v. State, 723 S.E.2d 876, 880–81 (Ga. 2012) (challenging that the state's public defender's system had a systematic breakdown which violated the defendant's speedy trial right).

361 *See supra* notes 17–24 and accompanying text.

362 Shelby Webb & John D. Harden, *Houston ISD Settles with Union over Controversial Teacher Evaluations*, HOUS. CHRON. (Oct. 12, 2017, 8:45 AM), https://www.chron.com/news/education/article/Houston-ISD-settles-with-union-over-teacher-12267893.php [https://perma.cc/Y4C6-8UCK].

363 Lecher, *supra* note 14.

364 AI NOW INST., LITIGATING ALGORITHMS: CHALLENGING GOVERNMENT USE OF ALGORITHMIC DECISION SYSTEMS 9 (2018), https://ainowinstitute.org/litigatingalgorithms.pdf [https://perma.cc/9EKU-2AHZ].

365 Settlement Agreement at 9–10, K.W. v. Armstrong, 180 F. Supp. 3d 703 (D. Idaho 2016), No. 1:12-cv-00022-BLW.

366 The risk of disproportionate liability has led some district courts to limit liability in cases of data breach. *Cf.* Storm v. Paytime, Inc., 90 F. Supp. 3d 359, 368 (M.D. Pa. 2015) ("[F]or a court to require companies to pay damages to thousands of customers, when there is yet to be a single case of identity theft

class action challenges to algorithmic governance techniques could be successful both in the sense of foreclosing the use of machine-learning tools in the absence of appropriate data and also catalyzing processes of analysis and reconstruction whereby the algorithm is not abandoned but improved. In this fashion, litigation supplies in part the necessary spur to check continuously for deviations from ground truth, to eliminate brittleness, and to account for distortions such as discrimination.

***

Regulation and litigation, as in many domains, are complementary partners in the catalysis of constitutional norms for the machine-learning state. Both are in their infancy now. There is almost no regulatory architecture in place at either the state or the federal level at the moment. There are a handful of suits challenging machine-learning tools. They provide useful proofs of concept. But neither the regulatory nor the litigation system is prepared, in sophistication or capacity, for the ongoing diffusion of algorithmic governance. As machine-learning tools spread across both the coercive, criminal justice state as well as its regulatory and welfare counterparts, there will be increasing cause to find an effectual regulatory architecture for the algorithmic state. This Part has begun that task by sketching the basic elements of the network of regulation and litigation necessary to ensuring that our algorithmic state is also a constitutional state.

CONCLUSION

Liberal constitutionalism entails a commitment to maintaining bounds on state power. That commitment is tested when "the technological and military character of governments and the productive relationships" of society change.[367] The "powerful and highly generalizable"[368] technology of machine learning poses a challenge to our constitutional system because it has the capability to transform the relationship between the state and its citizens.

I have suggested a suite of responses to that concern here. But more generally, I worry that new computational tools will

---

proven, strikes us as overzealous and unduly burdensome to businesses."). While an injunction might also impose costs on a public entity, it creates no perverse incentive to file socially negative value suits.

367 Shklar, *supra* note 1, at 24.

368 GREENFIELD, *supra* note 9.

tend to increase the capability of the state to analyze, predict, and control its subjects' behavior. They are also likely to decrease citizens' ability to understand and raises objections to coercive projections of state power. At the limit, the use of those technologies may cast doubt on the necessary conditions for the meaningful play of democratic control.

This potential asymmetry in power between the machine-learning state and its subjects (formerly citizens) presents a formidable challenge in the medium term. That challenge is most acute and most visible in China, where a range of surveillance and analytic technologies are deployed to suppress political dissent and leash ethnic and religious identity. But we should be under no illusions that the same technologies (and more) cannot find parallel uses in liberal democracies. Nor should we be under any illusion that steps explored here will on their own be sufficient to check the progress of a technocratic illiberalism. Far from it. Legal countermeasures of this ilk to the totalizing shadow of the state are always only adjuncts to larger, democratic efforts to keep the balance between state and citizen from capsizal. They will be effective only if conjoined with popular pressure, of the kind seen most recently in San Francisco's facial recognition ban, to check the machine-learning state when doing so remains within reach. It is the scale and passion of such public movements that will determine whether state algorithms comply with the rule of law, or whether instead they will be deployed to temper the democratic project.