PROCUREMENT AS POLICY: ADMINISTRATIVE PROCESS FOR MACHINE LEARNING

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

ABSTRACT

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

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

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

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

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

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

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

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

TABLE OF CONTENTS

I. INTRODUCTION

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

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

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

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

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

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

[4] *Id.*

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

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

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

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

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

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

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

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

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

[10] *Id.*

[11] *Id.* at 152.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A. THE ALGORITHMIC TURN IN GOVERNANCE

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

[24] Motor Vehicle Mfrs. Assn. v. State Farm Mut., 463 U.S. 29, 42–43 (1983).

[25] Mulligan & Bamberger, *supra* note 12.

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

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

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

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

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

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

[30] Brauneis & Goodman, *supra* note 6, at 107; *see* Coglianese & Lehr, *Regulating by Robot*, *supra* note 21, at 1161 (providing examples); Kimberly A. Houser & Debra Sanders, *The Use of Big Data Analytics by the IRS: Efficient Solutions or the End of Privacy as We Know It?*, 19 VAND. J. ENT. & TECH. L. 817 (2017) (discussing different ways the Internal Revenue Service uses data mining to solve the problem of tax noncompliance).

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

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

[31] *See* Marissa Cummings, *Federal Lawsuit Settled Between Houston's Teacher Union and HISD*, HOUS. PUB. MEDIA (Oct. 10, 2017), https://www.houstonpublicmedia.org/articles/news/2017/10/10/241724/federal-lawsuit-settled-between-houstons-teacher-union-and-hisd/ [https://perma.cc/JA6A-G3MW] (discussing artificial intelligence system used by the City of Houston, the Education Value-Added Assessment System (EVAAS), which made decisions about teacher evaluations, bonuses, and terminations based on variables including student's performance on prior standardized tests); *see also Settlement Agreement*, AFT.ORG (Oct. 2, 2017), https://www.aft.org/sites/default/files/settlementagreement_houston_100717.pdf [https://perma.cc/SC7W-XZEC] (agreeing that teachers would no longer be terminated based primarily on their EVAAS score).

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

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

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

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

[36] *Id.*

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

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

B. Challenges of Algorithmic Governance: Values in Technology Design

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[58] Burrell, *supra* note 51.

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

C. EXAMPLES: POLICY IN SYSTEM DESIGN

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

1. *Optimization Embeds Policy*

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

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

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

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

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

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

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

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

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

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

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

[65] *Id.*

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

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

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

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

[66] Davila, *supra* note 63.

[67] *Id.*

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

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

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

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

2. *Decisions About Target Variables Embed Policy*

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

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

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

[71] *Id.*

[72] *Id.*

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

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

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

3. *The Choice of Model Embeds Policy*

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

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

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

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

Id.

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

[77] *Id.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[83] Many scholars have offered excellent explanations and examples of data-related bias; these four are particularly rich and powerful: Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 677–94 (2016); Kate Crawford, *The Hidden Biases in Big Data*, HARV. BUS. REV. (Apr. 1, 2013), https://hbr.org/2013/04/the-hidden-biases-in-big-data [https://perma.cc/WC3X-CBJD]; and ROB KITCHIN, *Conceptualising Data*, *in* THE DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES & THEIR CONSEQUENCES 1–26 (2014).

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

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

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

5. *Decisions About Human-System Interactions Embed Policy*

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

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

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

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

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

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

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

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

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

[91] Wingate, *supra* note 33.

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

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

Id.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Id.

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

[101] *Id.* at 78.

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

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

III. BRINGING MACHINE-LEARNING SYSTEM DESIGN WITHIN ADMINISTRATIVE LAW

A. ADMINISTRATIVE PROCESS FOR MACHINE LEARNING DESIGN

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[112] *Id.* at 1304.

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

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

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

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

B. A Framework for Reasoned Decision Making About Machine Learning Design

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

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

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

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

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

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

[119] *Id.* at 496.

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

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

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

1. *Determining What System Choices Should Require Reasoned Decision Making*

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

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

a) Design Choices that Limit Future Agency Discretion

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

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

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

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

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

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

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

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

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

[124] *Id.*

[125] *Id.*

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

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

[128] *Id.*

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

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

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

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

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

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

[132] *Id.*

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

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

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

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

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

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

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

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

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

[136] *Id.* at 1111–13.

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

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

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

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

b) Normative Choices Between "Methods of Implementation"

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

[137] *Id.* at 1110–11.

[138] *Id.* at 1113.

[139] Id.

[140] Cmty. Nutrition Inst. v. Young, 818 F.2d 943 (D.C. Cir. 1987).

[141] 21 U.S.C. § 346.

[142] *Cmty. Nutrition Inst.*, 818 F.2d at 946–48 ("[T]his type of cabining of an agency's prosecutorial discretion can in fact rise to the level of a substantive, legislative rule."). That is exactly what has happened here.

[143] Hoctor v. U.S. Dep't of Agric., 82 F.3d 165, 171 (7th Cir. 1996).

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

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

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

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

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

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

[146] *Id.*

[147] *Id.*

[148] *Id.* at 170.

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

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

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

c) Application to Machine Learning Systems

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[163] Citron, *supra* note 19.

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

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

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

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

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

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

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

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

[166] *Id.* at 1304.

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

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

[169] *Id.* at 1305.

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

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

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

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

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

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

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

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

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

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

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

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

b. Element 2: Constraints on Policymaking Evolution

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

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

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

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

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

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

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

iii) The Challenge: Reintroducing Expert Justification for Agency Decisions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

IV. BUILDING ADMINISTRATIVE PROCESS FOR MACHINE LEARNING

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

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

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

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

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

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

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

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

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

While a range of agency processes might address the opacity of complex machine learning systems and account for the technocratic and democratic demands of reasoned governance,[201] this Part recommends elements for a framework of public machine learning adoption that satisfies both.

[196] *See supra* Section III.B.1 (discussing decision making by machine learning systems).

[197] *See* Mulligan & Bamberger, *supra* note 12 (discussing further the ways that values in technology change over time as technology is appropriated by users in new and unexpected ways, and how technology interactions with business models, organizational structures, and other technologies in ways that can transform its effects, use, and impact on values); Harry Surden, *Structural Rights in Privacy*, 60 SMU L. REV. 1605 (2007) (discussing ways that technology affects the "latent structural constraints" that work to protect values in addition to and in conjunction with legal measures).

[198] Kenneth A. Bamberger, *Provisional Precedent: Protecting Flexibility in Administrative Policymaking*, 77 N.Y.U. L. REV. 1272, 1280 (2002); *see* Matthew C. Stephenson, *Public Regulation of Private Enforcement: The Case for Expanding the Role of Administrative Agencies*, 91 VA. L. REV. 93, 139 (2005) ("Flexibility, like expertise, is often invoked to justify delegation of substantive policy choices to agencies.").

[199] *See, e.g.*, FCC v. Fox Television Stations, Inc., 556 U.S. 502 (2009) (applying arbitrary and capricious review to a change of agency policy applied in an adjudication); Motor Vehicle Mfrs. Assn. v. State Farm Mut., 463 U.S. 29, 30 (1983) (applying arbitrary and capricious review to a change in agency policy reached through rulemaking).

[200] Clark et al., *supra* note 50, at 463 (discussing how values tussles play out at design, redesign, configuration, and run time). This "developer" perspective on the ability and need to address values at every stage of the process is captured in the security adage, "Secure by Design, Secure by Default, Secure in Deployment." Steve Lipner, *The Trustworthy Computing Security Development Lifecycle*, 20TH ANN. IEEE COMPUTER SECURITY APPLICATIONS CONF. (2004).

[201] *See* TAXPAYER ADVOCATE SERV., *supra* note 60.

At its core is the reliance on centers of expertise—on the model of the USDS and the 18F "skunk works" team first developed by the Obama Administration—that develop and provide shared technical knowledge in ways that address expertise gaps across agencies while providing a systemic approach to the use of technology in government activity. We further identify different tools that such a centralized effort should employ, including algorithmic impact assessments, which not only involve deliberation about technical choices themselves, but also surface their policy implications publicly.

This framework uses two critical means to build on that visibility and foster public participation, political oversight, and informed agency engagement, during both system design and deployment. The first is institutional. It suggests that the adoption of processes that engage the public as policy is made through design. The second is technical. It suggests that reasoned agency deliberation about policy requires that machine learning systems adopted by governments reflect "contestable" design from the start—design that supports meaningful contestability throughout the system lifecycle by permitting an ongoing role for agency staff in shaping the policies embedded in systems.

Together, these tools focus on the development of expertise and the surfacing of politics while emphasizing judgment, coherence, efficiency, and transparency in setting administrative policy.

A. INFORMING AGENCY DELIBERATION WITH TECHNICAL EXPERTISE

1. *Reviewing Piecemeal Efforts*

In other works, we have argued that technology should be used to govern only when an agency has access to relevant technical expertise and the ability to consider a wide scope of public values that may be implicated in the use of technology to regulate.[202] We identified a range of options for acquiring relevant expertise, including agency hiring, drawing on the expertise of other agencies, and soliciting expertise from external stakeholders.[203]

Flavors of these alternative approaches to leveraging expertise are visible in some novel processes around algorithmic systems set up by select agencies and pending legislative proposals that seek to address both the potential for inherent bias and privacy risks.

Several jurisdictions at all levels have chosen to create public or quasi-public bodies to aid in the analysis of algorithmic-system adoption. At the

[202] Mulligan & Bamberger, *supra* note 12, at 759, 768–70. We also note that stakeholders must have the technical expertise to meaningfully participate and suggest alternative models for providing it. *Id.* at 775–76.

[203] *Id.* at 768–70.

local level, New York City passed an algorithmic accountability bill assigning a task force to examine the way city government agencies use algorithms.[204] New York's Automated Decision Systems Task Force, comprised of a cross-disciplinary group of city officials and outside experts,[205] is tasked with recommending a process for reviewing the city's use of automated decision systems to ensure equity and opportunity. The Task Force has held two public workshops to date.[206] While the law brings experts in to assist the city in developing review processes and conducting reviews, in recent testimony submitted to the New York City Council Committee on Technology, two Task Force members wrote:

> Task Force members have not been given any information about ADSs [automated decision systems) used by the City. To date, the City has not identified even a single system. Task Force members need to know about relevant systems used by the City to provide meaningful recommendations.[207]

They further reported that the Task Force had nevertheless made "meaningful progress in developing a methodology for eliciting relevant information about ADSs, using so-called "ADS Cards" that ask developers and operators to provide specific details about the system in question," but that the City forced them to abandon the project.[208]

At the state level, the Pennsylvania Commission on Sentencing[209] has brought expertise into the creation of their recidivism risk system in numerous ways. The Commission was tasked with creating a recidivism risk tool—initially paper-based but over time governed by software.[210] To

[204] *See* Int. No. 1696-A, Automated decision systems used by agencies (NYC 2018), http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0 [https://perma.cc/8AZR-DYB9].

[205] *Members*, NYC AUTOMATED DECISION SYSS. TASK FORCE, https://www1.nyc.gov/site/adstaskforce/members/members.page [https://perma.cc/R5VZ-LMQL] (last visited Oct. 10, 2019).

[206] *Past: Forum #2: Transparency*, NYC AUTOMATED DECISION SYSS. TASK FORCE, https://www1.nyc.gov/site/adstaskforce/events/events.page [https://perma.cc/LPR2-73JV] (last visited Oct. 10, 2019).

[207] *Testimony regarding Update on Local Law 49 of 2018 in Relation to Automated Decision Systems (ADS) Used by Agencies before NYC Council Comm. on Technology* (Apr. 4, 2019) (testimony of Julia Stoyanovich & Solon Barocas), https://dataresponsibly.github.io/documents/StoyanovichBarocas_April4,2019testimony.pdf [https://perma.cc/KY2W-M6PY].

[208] *Id.*

[209] The Commission was created in 1978 to develop and oversee the development of statewide guidelines to promote fairer and more uniform sentencing.

[210] Judicial Code and Prisons and Parole Code, 95 PA. STAT. ANN. §§ 42, 61 (2010) (providing for the adoption of a risk assessment tool).

develop these tools, the Commission has conducted its own research,[211] partnered with academic institutions to hold workshops,[212] commissioned academic research,[213] and most recently commissioned an independent evaluation by the Urban Institute.[214]

At the federal level, the Organ Procurement and Transplantation Network (OPTN) and Task Force on Organ Procurement and Transplantation, which were established to regulate the donation and allocation of organs through the network,[215] have a lengthy history of engaging experts in the design of their organ allocation systems.[216] The

[211] For overview of research, see *Research Overview of the Sentence Risk Assessment Instrument*, PA. COMMISSION ON SENT'G 4–5 (Oct. 2018), http://pcs.la.psu.edu/guidelines/proposed-risk-assessment-instrument/additional-information-about-the-proposed-sentence-risk-assessment-instrument/research-overview-of-the-sentence-risk-assessment-instrument-1/view; for reports to date, see PA. COMMISSION ON SENT'G, http://pcs.la.psu.edu/publications-and-research/risk-assessment [https://perma.cc/RH9P-CNRU] (last visited Aug. 1, 2019).

[212] *See Pennsylvania Criminal Justice Roundtable*, PENNSTATE CRIM. JUST. RES. CTR. (May 19–20, 2011), https://justicecenter.la.psu.edu/research/projects/pennsylvania-criminal-justice-roundtable [https://perma.cc/WFQ7-JF27] (convening state criminal justice policymakers and experts in offender risk assessment and sentencing, including faculty at the Penn State Criminal Justice Research Center).

[213] Matthew DeMichele & Julia Laskorunsky, *Sentencing Risk Assessment: A Follow-up Study of the Occurrence and Timing of Re-Arrest among Serious Offenders in Pennsylvania*, PA. COMMISSION ON SENT'G (May 2014), https://justicecenter.la.psu.edu/research/projects/files/PCS%20_Risk%20Assessment_Tool.pdf/view [https://perma.cc/PZ9T-TB5N] (analyzing the relationship between offender and case characteristics and likelihood of recidivism).

[214] *See* GEN. ASSEMB. COMM'N ON SENTENCING, PROPOSED SENTENCE RISK ASSESSMENT INSTRUMENT FOR 204 PA. CODE CHAPTER 305; RESPONSES TO PUBLIC COMMENTS; REQUEST FOR PROPOSALS; 48 Pa.B. 5445, at 2 (Sept. 1, 2018) (reporting that "in April 2018, following publication of a revised proposal, staff provided the Urban Institute with a complete set of files related to construction of the instrument (e.g., data, syntax, etc.) to begin the external review").

[215] National Organ Transplant Act, Pub. L. No. 98-507, 98 Stat. 2339 (1984) (establishing the Organ Procurement and Transplantation Network and banning organ sales). For a detailed legislative history covering the process and substantive considerations leading up to enactment, see Jed Adam Gross, *E. Pluribus UNOS: The National Organ Transplant Act and Its Postoperative Complications*, 8 YALE J. HEALTH POL'Y L. & ETHICS 145, 207–22 (2008).

[216] *See* Gross, *supra* note 215, at 228–30 (describing the NOTA Task Force on Organ Transplantation, which included "medical professionals, social and behavioral scientists, a legal scholar, an ethicist with a background in religious studies, representatives of the public and private insurance sectors, and representatives of the general public" who were tasked with developing recommendations for organ transplantation, including allocation policies); *id.* at 220 (discussing the requirement that procurement organizations include transplant professionals on their board of directors or advisory board); *see also* Organ Procurement and Transplantation Network, 42 C.F.R. § 121.12 (2019) (establishing the Advisory Committee

OPTN's Board of Directors, which develops policies to govern the operation of the OPTN, relies on numerous expert advisory committees.[217] Those committees engage in extensive fact-finding and deliberation.[218]

Since its inception, organ transplantation and allocation have been viewed as both highly technical and deeply political.[219] The algorithmic systems used to allocate organs are provided by the United Network for Organ Sharing (UNOS) under contract with the OPTN, which makes extensive information about the factors and weighting within allocation algorithms available.[220] OPTN/ONUS provides a central source of expertise around algorithmic allocation systems that governs the operation of all member transplant hospitals, organ procurement organizations, and histocompatibility labs in the United States. Thus, while numerous institutions are involved in procuring and transplanting organs, they benefit from a centralized set of expert resources. While bounded by legislation and overseen by the Department of Health & Human Services (HHS),[221] the OPTN/UNOS system for establishing policies to embed in the allocation

on Organ Transplantation, governed by the Federal Advisory Committee Act, to provide input on proposed OPTN policies and other matters); 42 C.F.R. § 121.3 (directing The OPTN Board of Directors to include "approximately 50% transplant surgeons or transplant physicians"); Mark D. Stegall et al., *Why Do We Have the Kidney Allocation System We Have Today? A History of the 2014 Kidney Allocation System*, 78 HUM. IMMUNOLOGY 4 (2017) (describing the deliberations and adoption of the kidney allocation system (KAS), including expert input, adopted in December 2014).

[217] For a list of current committees, see *Committees*, U.S. DEP'T HEALTH & HUM. SERVS., https://optn.transplant.hrsa.gov/members/committees [https://perma.cc/LWR6-6WWM] (last visited Aug. 1, 2019). Proposed policies OPTN wants to enforce, including allocation policies, must be provided to the Secretary of the Department of Health and Human Services sixty days prior to implementation, the Secretary must publish significant proposed policies in the Federal Register, and they are not enforceable until approved by the Secretary. Department of Health and Human Services. 42 C.F.R. § 121.4 (b)(2) (2019).

[218] For a description of a recent process for revising kidney allocations, see Mark D. Stegall et al., *Why Do We Have the Kidney Allocation System We Have Today? A History of the 2014 Kidney Allocation System*, HUM. IMMUNOLOGY 78.1 4, 4–8 (2017). For details on the allocation formula and discussion of use of a software system—Kidney-Pancreas Simulated Allocation Model (KPSAM)—to assess policy proposals, see Bhavna Chopra & Kalathail K. Sureshkumar, *Changing Organ Allocation Policy for Kidney Transplantation in the United States*, 5 WORLD J. TRANSPLANT 38 (2015).

[219] *See* Jed Adam Gross, *supra* note 215, at 182–83 (describing issues from blood pressure to distributive justice in congressional hearings).

[220] *See Organ Procurement and Transplantation Network Policies*, U.S. DEP'T HEALTH & HUM. SERVS., https://optn.transplant.hrsa.gov/governance/policies/ [https://perma.cc/JC5R-XL4R] (last visited Oct. 11, 2019) (detailing allocation rules and weighting for various organs).

[221] The Final Rule requires organ allocation to be based on sound medical judgment; make the best use of donated organs; avoid wasting organs and futile transplants; and promote patient access to transplants. 42 C.F.R. § 121.8 (2019).

algorithms is largely driven by experts enlisted through committees. Those committees are highly specialized, focusing on the specific factors that influence transplant success and considerations of just allocation with respect to particular organs.

The Taxpayer Advocate's office within the IRS offers an alternative model focused on agency development of internal expertise. While not set up specifically to address algorithmic systems, the Taxpayer Advocate has played a key role in identifying the problem with embedded policies in algorithmic systems.[222] Although the Advocate has not been fully successful in ensuring that embedded policies are vetted internally, consistent with other agency review and approval processes for all written policy, procedures, and guidance before issuance and publication, they have drawn agency attention to problematic aspects of systems, some of which have been reformed.[223]

Finally, a bill pending in California takes a different outward-looking approach. Rather than bringing expertise directly into government work, it would leverage the expertise of firms that provide AI-based products and services to public agencies. The bill would require firms to provide information about data curation and processing, as well as bias mitigation strategies, in their contracts with public agencies.[224] Requirements such as this could harness the knowledge and expertise of those close to the problem toward public ends—allowing the experts to identify and address problems, but opening them up for regulator and public scrutiny.[225]

Over the past year, engineers and other employees at firms that build machine learning systems have been increasingly active in objecting to their

[222] TAXPAYER ADVOCATE SERV., *supra* note 60 (arguing that "automated systems and software applications require transparency, and employee guidance embedded in systems must be reviewed and continually analyzed for proper application" but noting that "policy guidance embedded in systems is neither reviewed internally nor published externally").

[223] *Id.*

[224] Artificial Intelligence: Reporting, S.B. 444, 2019 Legis., Reg. Sess. (Cal. 2019), was a state senate bill in California introduced by Senator Umberg (D) on February 21, 2019 to establish that a contractor, vendor, or qualifying business shall maintain a written record of the data used relating to any use of artificial intelligence for the delivery of a product or service to a public entity. The records shall include all the following information: (1) the purpose of the data; (2) a description of the categories of the data; (3) the source of the data; (4) the demographics or information related to a characteristic listed in subdivision (a) of Section 11135 of the Government Code that is used as a source of input data for the creation of the artificial intelligence system. The business shall disclose to public entities that it relies on AI, information about the data, and "its internal policies for how bias in the artificial intelligence system is identified and mitigated." *Id.* at § 3505(c)(3).

[225] *See generally* Kaminsky, *supra* note 22, at 26–39 (calling for collaborative governance of algorithms used in the private sector).

use in specific contexts or toward specific ends.[226] Providing engineers and other technical professionals with external justifications for considering the implications of their work on privacy, discrimination, and other substantive values can legitimize these nascent expressions of concern among technical professionals, encourage internal policing,[227] and provide a platform to support collaborative work across institutions on issues such as bias mitigation strategies.

Each of these approaches has their advantages. Given the current level of uncertainty about how best to mitigate bias, leveraging the deep knowledge of the people and institutions close to the problem—as the OPTN and California proposal do in different ways—while exposing them to external review taps into the relevant expertise while fostering attention and accountability to a broader set of values.[228] The Pennsylvania Sentencing Commission, OPTN, and the California proposals lay different internal processes out for the public. OPTN and the Pennsylvania Sentencing Commission provide more detailed information about the properties of the algorithmic systems themselves. Requiring decision makers to explain themselves to outsiders with different perspectives on what policies ought to be embedded, or what bias mitigation strategies should be taken, can push those developing algorithmic systems to engage in more critical technical practice.[229] This approach may yield processes that make agencies, managers, or even individual engineers accountable for considering the impact of algorithmic systems from perspectives beyond speed and task performance, fostering a greater sense of responsibility for the outcomes they produce in the world.

In contrast to the Sentencing Commission and OPTN, the New York City Task Force's mandate cuts across numerous agencies. While the factors contributing to its slow progress are not fully clear, one factor may be that focusing on algorithmic systems writ large, rather than algorithmic systems

[226] Meredith Whittaker et al., *AI Now Report 2018*, AI NOW INST. 40–42 (Dec. 2018), https://ainowinstitute.org/AI_Now_2018_Report.pdf [https://perma.cc/7WER-78L4] (summarizing host of employee actions at various companies to stop particular AI projects and system uses).

[227] Such approaches could bolster algorithmic whistleblowing by employees envisioned by Sonia Katyal. Katyal, *supra* note 20.

[228] *See* KENNETH A. BAMBERGER & DEIRDRE K. MULLIGAN, PRIVACY ON THE GROUND: DRIVING CORPORATE BEHAVIOR IN THE UNITED STATES AND EUROPE 190 (2015) (describing importance of exposing business practices to scrutiny) [hereinafter PRIVACY ON THE GROUND]; *see generally* Kenneth A. Bamberger, *Regulation as Delegation: Private Firms, Decisionmaking, and Accountability in the Administrative State*, 56 DUKE L.J. 377, 445–46 (2006).

[229] Philip E. Tetlock, *Accountability: The Neglected Social Context of Judgment and Choice*, 7 RES. ORGANIZATIONAL BEHAV. 297, 314–21 (1985) (reviewing research evidence).

within a particular domain and with participation by substantive experts in that domain, may be perceived as more threatening or less beneficial.

2. *A Paradigm for Expert Decision Making*

However, these piecemeal attempts to inform case-by-case deliberation with internally- or externally-derived technical expertise about methodologies under consideration offer only a proverbial finger in the dike[230] against the oncoming flood of government usage of machine learning. While helpful, they are fragmentary, costly, and time-consuming. Furthermore, they often (with the possible exception of the New York City Task Force) fail to address the expertise gap across agencies by providing easily deployable, coordinated, and regularized policies and methods governing, and processes for assessing, algorithmic governance tools.

Indeed, centralization of expertise may be particularly important for attending to embedded policies in algorithmic systems. Organ donation is an area where the need for allocation strategies and their deeply political and technical nature were recognized and accounted for up front. But in many areas, algorithmic systems are replacing or aiding decision making that was conceived of and practiced in a more clinical way—relying on human judgment informed by expertise gained through education and training, refined through tacit knowledge and intuition developed experientially through practice, and perhaps discussion and feedback with others—and therefore lacks the level of formalization found in organ donation or even recidivism risk. To the extent algorithmic systems are being introduced in areas where clinical decision-making methods reign, the need for relevant technical expertise in addition to domain-specific expertise will be high; however, internal agency capacity to provide it may be low.

a) The Institutional Paradigm: USDS and the 18F "Skunk Works"

In this vein, efforts begun during the Obama Administration to create centers-of-expertise or shared technical knowledge offer a paradigm for a more systemic approach to the use of technology in government activity. This work has continued with the Office of Federal Procurement Policy (OFPP),[231] which, in conjunction with the USDS, created a Digital IT Acquisition Program (DITAP) to help contracting professionals gain the

[230] *See* MARY MAPES DODGE, HANS BRINKER OR THE SILVER SKATES: A STORY OF LIFE IN HOLLAND 105–09 (recounting a tale about a Dutch boy who saves his country by putting his finger in a leaking dike).

[231] OFPP was created by Congress to provide overall direction on "Government-wide procurement policies, regulations, procedures, and forms"; and to "promote economy, efficiency, and effectiveness" in procurement. 41 U.S.C. § 1101(b) (2012).

"expertise needed to support the delivery of digital information (i.e., data or content) and transactional services (e.g., online forms and benefits applications) across a variety of platforms, devices, and delivery mechanisms (e.g., websites, mobile applications, and social media)."[232]

President Obama created a centralized pool of technical expertise, USDS, housed within the Executive Office of the President of the United States. USDS is a transitory pool of experts in design, engineering, or product management brought in from outside government for "tours" of service.[233] USDS provides consultation services to federal agencies on information technology and guidance. For example, to develop greater technical expertise among the contracting professionals distributed throughout federal agencies, USDS created the Digital Services Playbook to propagate best practices from both the private and public sector across federal agencies. The Playbook is accompanied by the TechFAR Handbook, which provides guidance to agencies on how to comply with Federal Acquisition Regulations while using the Digital Services Playbook. Similarly, 18F—sometimes described as a skunk works[234] project for government—is an office within the General Services Administration (GSA) that collaborates with other agencies to fix technical problems, build products, and improve how government serves the public through technology.[235] They offer agencies access to expert teams of designers, software engineers, strategists, and product managers skilled in user-centered development, and other design and acquisition expertise.

But in addition to providing experts to government agencies on a task-specific model, 18F also develops guides on topics such as accessibility, agile

[232] Memorandum from Lesley A. Field, Deputy Adm'r, Office of Mgmt. & Budget, on Establishment of Federal Acquisition Certification in Contracting Core-Plus Specialization in Digital Services (FAC-C-DS) to Chief Acquisition Officers & Senior Procurement Execs. A-1 (May 18, 2018), https://techfarhub.cio.gov/assets/files/FAC_C_Digital_Services_5-18-18.pdf [https://perma.cc/75MU-WWS9].

[233] *How We Work*, U.S. DIGITAL SERV., https://www.usds.gov/how-we-work [https://perma.cc/QZY3-YV6C] (last visited Oct. 3, 2019).

[234] *See* Dave Zvenyach, *Joining 18F*, V. DAVID ZVENYACH'S BLOGS (Jan. 20, 2015), https://esq.io/blog/posts/joining-18f/ [https://perma.cc/VM9T-F4C8] (describing 18F as "a modern-day digital skunk works . . . [housing] some of the best and brightest developers in America, building applications that agencies need in a modern, experimental, and explicitly iterative manner"); *accord* Jason Bloomberg, *Digital Influencer Jez Humble: DevOps For 'Big Hairy Enterprises'*, FORBES (Mar. 31, 2016, 9:54 AM), https://www.forbes.com/sites/jasonbloomberg/2016/03/31/digital-influencer-jez-humble-devops-for-big-hairy-enterprises/#2f603f054a21 [https://perma.cc/8B3G-C8R4] (describing 18F as "a skunkworks-like team of designers, developers, and product specialists").

[235] *About*, 18F, https://18f.gsa.gov/about/#our-team [https://perma.cc/2DT4-K9FQ] (last visited Oct. 3, 2019).

development, and design methods to assist federal agencies.[236] An example of such guidance documents is the U.S. Web Design Standards (USWDS),[237] a joint product of 18F and USDS.[238] The USWDS provide guidance, templates, and models for developers and designers; it covers design including accessibility, front end and back end coding, and provides code and performance guidelines.[239] The mix of centralized expert staff who can be deployed for limited periods of time to assist agencies with complex technology projects and detailed guidance documents is appealing in the context of algorithmic systems.

Coordinated expert input through shared resources in the vein of the USWDS should frame agency decision making around algorithmic systems, as well as consulting services to provide hands-on assistance to agencies on an as-needed basis. The New York City Task Force was an effort to develop a methodology for eliciting relevant information from developers and operators of systems through targeted questions. It builds on research in the FAT* community identifying specific information necessary to understand appropriate uses of and potential biases in systems.[240] The recently introduced Algorithmic Accountability Act of 2019,[241] while aimed at the private sector rather than the public sector, takes a similar approach by requiring automated decision system, data protection impact assessments, and regulations promulgated by the Federal Trade Commission to guide such assessments. A similar mix of standardized questions, guidance, and localized

[236] *Guides*, 18F, https://18f.gsa.gov/guides/ [https://perma.cc/N2SD-QY7B] (last visited Oct. 3, 2019).

[237] U.S. WEB DESIGN SYS., https://designsystem.digital.gov/ [https://perma.cc/2V2Y-K2HN] (last visited Oct. 3, 2019).

[238] Mollie Ruskin et al., *Introducing the U.S. Web Design Standards*, 18F (Sept. 28, 2015), https://18f.gsa.gov/2015/09/28/web-design-standards/ [https://perma.cc/X33R-P43C].

[239] U.S. WEB DESIGN SYS., *supra* note 237.

[240] *See* Margaret Mitchell et al., *Model Cards for Model Reporting*, PROC. CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 220 (2019) (proposing information about what the use context model was designed for, model performance benchmarked across different groups, and processes of validation and accompany models); *see also* Galen Harrison et al., *Towards Supporting and Documenting Algorithmic Fairness in the Data Science Workflow*, WORKSHOP ON TECH. & CONSUMER PROTECTION (May 23, 2019), https://www.ieee-security.org/TC/SPW2019/ConPro/papers/harrison-conpro19.pdf [https://perma.cc/VPZ9-B4UN] (proposing documentation and visualization of algorithms in data science processes); *see generally* Timnit Gebru et al., *Datasheets for Datasets*, ARXIV (Apr. 16, 2019), https://arxiv.org/pdf/1803.09010.pdf [https://perma.cc/82DT-G93P] (proposing information about attributes of datasets that should be documented and shared). These efforts resemble the efforts in the reproducible research community to provide information about data, code, computational steps, software environment, etc. *See* Victoria Stodden et al., *Enhancing Reproducibility for Computational Methods*, 354 SCIENCE 1240 (2016).

[241] S. 1108, 116th Cong. (2019); H.R. 2231, 116th Cong. (2019) (providing the companion House bill).

assessment is found in the privacy area, where Congress required administrative agencies to conduct privacy impact assessments[242] but authorized the Office of Management and Budget (OMB) to issue detailed guidance for agency implementation.[243]

The mix of expertise offered by 18F and USDS, and their track record of providing effective guidance and leadership, provides the most compelling model to begin developing methods and tools for cross-agency efforts to identify and reason about embedded policies in algorithmic decision-making systems.

b) Models to Inform the Centralized Process

Existing legal frameworks addressing the use of predictive models in the area of credit and employment offer guidance regarding the possible content of centralized processes. For example, under the Equal Credit Opportunity Act (ECOA), in order to use age[244] as a predictive factor in granting credit, creditors must use an "empirically derived, demonstrably and statistically sound, credit scoring system."[245] To meet these criteria, the system must be:

> (i) [b]ased on data that are derived from an empirical comparison of sample groups or the population of creditworthy and non-creditworthy applicants who applied for credit within a reasonable preceding period of time;
>
> (ii) [d]eveloped for the purpose of evaluating the creditworthiness of applicants with respect to the legitimate business interests of the

[242] E-Government Act of 2002, Pub. L. No. 107-347, 116 Stat. 2899, 2921 (2002) (requiring agencies to conduct a privacy impact assessment before "developing or procuring information technology that collects, maintains, or disseminates information that is in an identifiable form").

[243] *See* Memorandum from Joshua B. Bolten, Dir., Office of Mgmt. & Budget, on OMB Guidance for Implementing the Privacy Provisions of the E-Government Act of 2002 to Heads of Exec. Dep'ts & Agencies (Sept. 26, 2003), https://obamawhitehouse.archives.gov/omb/memoranda_m03-22/ [https://perma.cc/39ZE-U2TC].

[244] While these requirements only apply to the use of credit scoring systems that use age as a predictive factor in granting credit, regulator comments suggest that in practice, statistical validation of credit scoring systems is a key tool that the Federal Trade Commission and Consumer Financial Protection Board use to assess compliance with ECOA and that regulated entities meet these guidelines to manage risk. *See Credit Scoring: Testimony Before the U.S. House Subcomm. on Financial Institutions & Consumer Credit*, 111th Cong. (Mar. 24, 2010), https://www.federalreserve.gov/newsevents/testimony/braunstein20100323a.htm [https://perma.cc/ZU98-N33W] (testimony of Sandra F. Braunstein, Dir., Div. of Consumer & Cmty. Affairs, Fed. Reserve).

[245] *See* 12 C.F.R. § 1002.2(p) (defining empirically-derived and other credit scoring systems); 12 C.F.R. app. I § 1002.2(p) (2019) ("1. . . . The definition under §§ 1002.2(p)(1)(i) through (iv) sets the criteria that a credit system must meet in order to use age as a predictive factor.").

> creditor utilizing the system (including, but not limited to, minimizing bad debt losses and operating expenses in accordance with the creditor's business judgment);
>
> (iii) [d]eveloped and validated using accepted statistical principles and methodology; and
>
> (iv) [p]eriodically revalidated by the use of appropriate statistical principles and methodology and adjusted as necessary to maintain predictive ability.[246]

While designed to regulate credit granting, these criteria, along with the research on aspects of models and data sets necessary to determine appropriate and fair uses, provide a useful starting point for thinking about data and models.

The Uniform Guidelines on Employee Selection Procedures (Uniform Guidelines)[247] issued in 1978 by the Equal Employment Opportunity Commission, provides another set of criteria that might be useful in the context of developing guidance on algorithmic systems. These guidelines are designed "to assist employers, labor organizations, employment agencies, and licensing and certification boards to comply with requirements of Federal law prohibiting employment practices which discriminate on grounds of race, color, religion, sex, and national origin."[248] The Uniform Guidelines specify that selection procedures having an adverse impact on these protected characteristics will be prohibited:

> The use of any selection procedure which has an adverse impact on the hiring, promotion, or other employment or membership opportunities of members of any race, sex, or ethnic group will be considered to be discriminatory and inconsistent with these guidelines, unless the procedure has been validated in accordance with these guidelines, or the provisions of section 6 below are satisfied.[249]

The guidelines thus provide a framework for determining the proper use of tests and other selection procedures. Moreover, while they do not require employers to conduct validity studies where no adverse impact results, they draw attention to the potential disparate impacts of selection procedures by providing guidance on how to construct and validate them—encouraging more thoughtful technical choices. For example, the Uniform Guidelines state that "where two or more selection procedures are available which serve

[246] 12 C.F.R. § 1002.2(p), n.217.

[247] 29 C.F.R. § 1607 (2019).

[248] 29 C.F.R. § 1607.1.

[249] 29 C.F.R. § 1607.3.

the user's legitimate interest . . . and which are substantially equally valid for a given purpose, the user should use the procedure which has been demonstrated to have the lesser adverse impact."[250] While these guidelines already apply to the employment practices of agencies, they are equally applicable to the development of algorithmic systems more generally and could easily be adapted for such purposes. Other elements of the Uniform Guidelines are similarly relevant, such as the discussion of criterion-related, content, and construct validity studies to assess selection procedures.[251] In particular, the Uniform Guidelines clarify that the validity of a selection procedure can be validated by empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance; by data showing that the content of the selection procedure is representative of important aspects of performance on the job; or through a construct validity study that consists of data showing that the procedure measures the degree to which candidates have identifiable characteristics which have been determined to be important in successful performance in the job.[252] In particular, the distinction between criterion-related validity and construct validity could be used to clarify different ways in which algorithmic systems could be reviewed. Although the Uniform Guidelines were issued thirty years ago, their focus on the *impact* of selection procedures—procedures used to predict how an applicant will do at a job—continues to make them relevant in the context of automated decisions today.

In conjunction with the growing set of guidance documents, toolkits, and other methods for analyzing algorithmic systems, these existing government regulations could provide a starting point for developing a suite of guidance documents and methods to standardize the questions and processes federal agencies use to assess and design algorithmic systems. With the availability of appropriate experts at 18F and USDS, this could provide a flexible, on-demand set of tools and personnel to fill the expertise gap plaguing federal agencies.

B. INFUSING AGENCY DELIBERATION WITH POLITICAL VISIBILITY

1. *Impact Assessments: Bridging Technocracy and Democracy in Agency Deliberation*

Algorithmic impact assessments provide a critical tool to bridge the technocratic and democratic elements of deliberation.[253] Such tools not only

[250] 29 C.F.R. § 1607.3.B.

[251] *See, e.g.*, 29 C.F.R. § 1607.5.B.

[252] *Id.*

[253] *See* Reisman et al., *supra* note 22; Selbst, *supra* note 22.

enable and trigger agency deliberation about the technical aspects of system design, but surface the political implications of those choices, offering an important prerequisite for reasoned consideration of the policies they embed by agency staff, the public, and the political branches. They thus bridge the dual deliberation requirements of substantive expertise and political visibility.

More specifically, algorithmic impact assessments are boundary negotiation objects—objects used to "record, organize, explore and share ideas; introduce concepts and techniques; create alliances; create a venue for the exchange of information; augment brokering activities; and create shared understanding."[254] If publicly disclosed, algorithmic impact assessments mediate between experts and the public, providing a common reference point, but importantly do not reflect a common understanding or consensus across these groups.[255] A good impact assessment provides a tool for exploring the problem space, helping the public collectively consider the points of policy within a machine learning system. They also allow the public and expert communities to in effect argue about the boundary between science and policy; in this way, they facilitate negotiation not just across the boundary, but about the location of it.

It is no accident that many of the "arbitrary and capricious" cases finding that agency action fails to reflect appropriate deliberation about relevant analytic techniques (models, assumptions, the use of data, and choices between false-negatives and false-positives) arise in the review of either cost-benefit or environmental-impact statements—two types of impact assessments well-enshrined in administrative law. Such tools are instrumental in "making bureaucracies think,"[256] and "take a hard look at the potential . . . consequences of their actions."[257] While they do not mandate substantive

[254] Matthew J. Bietz & Charlotte P. Lee, *Collaboration in Metagenomics: Sequence Databases and the Organization of Scientific Work*, PROC. 11TH EUR. CONF. ON COMPUTER SUPPORTED COOPERATIVE WORK 243, 247.

[255] (2009) Privacy regulators and scholars generally advocate for publishing privacy impact assessments to support "contestation and public debate." Jennifer Stoddart, *Auditing Privacy Impact Assessments: The Canadian Experience*, *in* PRIVACY IMPACT ASSESSMENT 453–54 (David Wright & Paul De Hert eds., 2012) (describing various regulators' positions on publication). Unfortunately, proposed bills in the United States do not require publication. *See* The Algorithmic Accountability Act of 2019, S. 1108, 116th Cong. (2019); H.R. 2231, 116th Cong. (2019); *see also* Margot E. Kaminski & Andrew D. Selbst, *The Legislation That Targets the Racist Impacts of Tech*, N.Y. TIMES (May 7, 2019), https://www.nytimes.com/2019/05/07/opinion/tech-racism-algorithms.html [https://perma.cc/P6YV-43AV] (critiquing the lack of a publication requirement).

[256] SERGE TAYLOR, MAKING BUREAUCRACIES THINK: THE ENVIRONMENTAL IMPACT STATEMENT STRATEGY OF ADMINISTRATIVE REFORM 251 (1984).

[257] *The National Environmental Policy Act: A Study of Its Effectiveness After Twenty-Five Years*, COUNCIL ON ENVTL. QUALITY iii (Jan. 1997), https://ceq.doe.gov/docs/ceq-publications/nepa25fn.pdf [https://perma.cc/4KJH-4VK5] (discussing the National

outcomes, they force administrative agencies to turn their analytic capacity towards particular issues, and require explicit and publicly reviewable identification, recognition, and explanation of their choices about them.

Such requirements are especially important in surfacing issues in contexts with which government actors are unfamiliar, or where the issues addressed are "orthogonal to" or even "in tension with, an agency's primary mission."[258] Our research has demonstrated the particular importance of such impact assessments when agencies engage with the use of technology, given problems of technical illiteracy, design opacity, and the phenomena by which the modality "hide[s] the political nature of its design,"[259] rendering policy implications invisible and making choices seem fixed, natural, and incontestable. In particular, the requirement imposed by the E-Government Act of 2002,[260] that agencies complete privacy impact assessments (PIAs) when developing or procuring information technology systems that include personally identifiable information, has forced agencies to address important privacy implications of systems intended to promote a range of public aims—implications that remained unnoticed and unaddressed (with important and expensive security consequences) where such requirements were not satisfied robustly.[261]

When agencies adopt technology, we have argued accordingly, the choices that impact legal rights must be addressed through the use of impact assessment tools.[262] On the one hand, those tools "create different frameworks and bring new considerations to bear in agency actions," as well as bridge the gulf between the substantive domain expertise of agency staff and the frameworks and knowledge of outside experts.[263] On the other, they "facilitate participation by issue experts and by stakeholders who might otherwise be unaware of relevant risks and technological alternatives."[264]

Environmental Policy Act's "success" in making federal agencies take a "hard look" at the potential environmental consequences of their actions).

[258] Kenneth A. Bamberger & Deirdre K. Mulligan, *Privacy Decisionmaking in Administrative Agencies*, 75 U. CHI. L. REV. 75, 83 (2008).

[259] Mulligan & Bamberger, *supra* note 12, at 778.

[260] E-Government Act of 2002, Pub. L. No. 107-347, 116 Stat. 2899, 2921 (2002).

[261] *See* Kenneth A. Bamberger & Deirdre K. Mulligan, *PIA Requirements and Privacy Decision-Making in US Government Agencies*, *in* PRIVACY IMPACT ASSESSMENT 225–50 (David Wright & Paul De Hert eds., 2012) (comparing DHS's responsible adoption of RFID technology with that of the Department of State, which failed to discuss technical aspects of the program, alternative technologies, and risks); Bamberger & Mulligan, *supra* note 258 (discussing the same).

[262] Mulligan & Bamberger, *supra* note 12, at 764–66, 780 (arguing that when agencies govern "by design," they should use values-impact assessment tools, such as a "human rights impact assessment").

[263] *Id.* at 765.

[264] *Id.*

To have the desired ameliorative effect, experts must be involved in conducting impact assessments[265] and the outputs must be available to the public.[266] Current and proposed efforts to require government agencies to conduct impacts of surveillance and algorithmic systems too often emphasize process while shorting or overlooking expertise.[267] There is an ongoing rise in chief privacy officers and other privacy staff, and more recently a move to redefine their roles to provide them greater latitude to address harms that may flow from data analysis and applications—reflected in new titles such as chief data governance officer or information steward.[268] More recently, growing concerns with AI—ranging from job displacement to bias to military applications—have ushered in a new set of professional staff with titles such as chief ethics officer.[269] More significantly, a major shift in privacy regulation in the European Union was to require data protection officers. The emphasis on professionals reflects the growing understanding that particular expertise is necessary to fully use tools such as impact assessments.

2. *Other Political Visibility-Enhancing Processes*

The impact assessments described above provide foundational tools for spanning boundaries between expert communities, policy makers, and the

[265] Bamberger & Mulligan, *supra* note 258, at 104 (concluding that optimal use of privacy impact assessment turned on expert inter-disciplinary staff).

[266] While different countries have taken different positions on whether privacy impact assessments or summaries ought to be shared with the public, there is a general preference for doing so due to the recognition that doing so can maintain public trust and confidence in systems and organizations. *See* David Wright & Paul De Hert, *Introduction to Privacy Impact Assessment*, *in* PRIVACY IMPACT ASSESSMENT 3–32 (David Wright & Paul De Hert eds., 2012); *see also* Elin Palm & Sven Ove Hansson, *The Case for Ethical Technology Assessment (eTA)*, 73 TECHNOLOGICAL FORECASTING & SOC. CHANGE 543, 547 (2006) (arguing for publication because "[i]t would be delusive to believe that technology developers are conscious of all the effects of their products[; i]n many cases, negative side effects come as a surprise to technology developers themselves").

[267] *See* OAKLAND, CAL., MUN. CODE §§ 9.64.010–9.64.070 (2018) (establishing requirement for impact assessments but not requiring new staff); *accord* SANTA CLARA CITY, CAL., ORDINANCE CODE § A40 (2016).

[268] *See, e.g.*, Molly Hulefeld, *What is a chief data ethics officer anyway*, IAPP PRIVACY ADVISORY (Nov. 27, 2018), https://iapp.org/news/a/making-way-for-the-rise-of-the-chief-data-ethics-officer/ [https://perma.cc/Q39K-WDUV] ("Acxiom renamed its privacy program to become the[']data ethics, governance, protection and privacy program.' "); *Mastercard Names JoAnn Stonier Chief Data Officer*, MASTERCARD (Feb. 8, 2018), https://newsroom.mastercard.com/press-releases/mastercard-names-joann-stonier-chief-data-officer/ [https://perma.cc/BT43-QVX6] (announcing that JoAnn Stonier was moving from chief information governance and chief privacy officer to chief data officer, a new position designed to affirm the company's commitment to data protection).

[269] *See, e.g.*, *Rise of The Chief Ethics Officer*, FORBES (Mar. 27, 2019), https://www.forbes.com/sites/insights-intelai/2019/03/27/rise-of-the-chief-ethics-officer/#5e67f2ba5aba [https://perma.cc/8P84-HFYR].

general public. By distilling the policy-relevant choices in technical design, impact assessments surface issues for political consideration.

However, the question of what aspects of a technical system are political is itself a value judgment, and who decides is itself a political matter. Despite the best efforts of experts,[270] the political and the technical defy clean separation.[271] The complexity and density of technical and scientific matters can create barriers to broader participation in policy debates. This is surely true with respect to machine learning systems, where the complex interaction of design choices, data, and interfaces can produce clearly political outcomes yet be shielded from public scrutiny. Thus, agencies must further employ a broader set of processes to publicize system politics and elicit the public participation essential for legitimate administrative decision making.

The COMPAS debates provide a glimpse of the politics of expertise in action in machine learning systems. After journalists at ProPublica exposed the different false positive and false negative rates for black and white defendants, Northpointe, the developer of COMPAS, defended the system because it was "equally accurate for blacks and whites," asserted its status as expert on the matter, and dismissed ProPublica's analysis as wrong because they failed to "account the different base rates of recidivism for blacks and whites."[272] Their rejoinder attempts to remove a legitimate political question—should models of fairness account for different base rates—from the public discussion by framing the question as one of whether to account for objective facts. Academics later pointed out that the differing perspectives on how to conceive of fairness espoused by Northpointe and ProPublica were mathematically incompatible, yet both defensible and possibly mutually desirable.[273] Here, as in many other instances, the technical

[270] Jasanoff has documented how experts in science policy constantly attempted to demarcate the politics from the scientific in their practice in an effort to preserve claims of objectivity. *See* Sheila S. Jasanoff, *Contested Boundaries in Policy-Relevant Science*, 17 SOC. STUDS. SCI. 195, 199 (1987) ("To shore up their claims to cognitive authority, scientists have to impose their own boundaries between science and policy.").

[271] *See generally* STATES OF KNOWLEDGE: THE CO-PRODUCTION OF SCIENCE AND SOCIAL ORDER (Sheila Jasanoff ed., 2004) (presenting a collection of essays exploring the ways in which our methods of understanding and reasoning about the world and choices about how to live in the world are interdependent).

[272] William Dieterich et al., *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, NORTHPOINTE 9 (July 8, 2016), http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf [https://perma.cc/WQQ5-SEMC].

[273] *See* Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear*, WASH. POST (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.ce67b2c3fa95 [https://perma.cc/5TDR-E4MQ]; *see also* Jon Kleinberg et al., *Inherent Trade-*

and the political are entangled. Yet the important question about what metric should be used to support fair risk assessments appears to have escaped careful scrutiny by the public agency.

Fostering public engagement requires processes at a variety of levels, and times, to surface the politically salient questions latent in the design and use of machine learning systems. As an initial matter, public engagement would *both* be primed at the abstract level—i.e., the public would have a general understanding of the important policy choices entangled with system design—*and* invited to participate in discourse with respect to specific systems.

a) Fostering Ongoing Public Engagement Through Agenda-Setting

Today, the politics of machine learning systems are an object of ongoing public scrutiny. The "black boxes" are being aired out, and centers of science policy have participated in opening up algorithmic systems to scrutiny. During the Obama Administration, the Office of Science and Technology Policy wrote several reports exploring the politics of the design and use of big data and artificial intelligence.[274] In particular, the reports detailed the biases that can result from training data and model design.[275] The sustained attention to the political issues embedded in technical systems was particularly important given President Obama's efforts to strengthen "America's role as the world's engine of scientific discovery and technological innovation" and use technology to improve service delivery

Offs In The Fair Determination of Risk Scores, ARXIV (Nov. 17, 2016), https://arxiv.org/pdf/1609.05807.pdf [https://perma.cc/MR4S-9W8W].

[274] *See, e.g.*, *Big Data: Seizing Opportunities, Preserving Values*, EXECUTIVE OFF. PRESIDENT (May 2014), https://hsdl.org/?view&did=752636 [https://perma.cc/AV9Y-L4FR]; *Big Data and Differential Processing*, EXECUTIVE OFF. PRESIDENT (Feb. 2015), https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/docs/Big_Data_Report_Nonembargo_v2.pdf [https://perma.cc/Z9AL-48C7]; *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, EXECUTIVE OFF. PRESIDENT (May 2016), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf [https://perma.cc/VJ3S-HW3G]; Ed Felten & Terah Lyons, *The Administration's Report on the Future of Artificial Intelligence*, WHITE HOUSE BLOG (Oct. 12, 2016, 6:02 AM), https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence [https://perma.cc/X4EQ-CFWT]; *see also* John P. Holdren & Megan Smith, *Cabinet Exit Memo*, OFF. SCI. & TECH. POL'Y, EXECUTIVE OFF. PRESIDENT (Jan. 5, 2017), https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_exit_memo_final.pdf [https://perma.cc/FN5W-7S6B].

[275] *See Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, *supra* note 274.

and governance across the public sector.[276] The Federal Trade Commission conducted workshops and reports that similarly focused attention on the policy implications of algorithmic systems, but with attention to its implications in the marketplace rather than the civic sector.[277] These reports focused agencies and the public on the politics of algorithmic choices, and sent a strong political signal to agencies that technological design ought to be scrutinized as policy.

Agenda-setting activities such as these White House reports, other Administration strategy documents that address the ethical, legal, and social aspects of artificial intelligence,[278] and studies and reports by organizations such as the National Academy of Science,[279] throw open the doors to ongoing public engagement through plain language, case studies, context, and explicit identification of the policy judgments that warrant public engagement along with technical expertise. They also encourage agencies to surface questions about specific technical system designs for public scrutiny.

b) Fostering Public Engagement on Specific Systems

Broader participation in the adoption and design of specific systems poses some practical challenges. As discussed in Part II, unlike regulations which agencies write themselves, the software code of machine learning systems is typically authored and owned by companies. Empirical work suggests that agencies routinely have little impact on the design of such systems, largely procuring them "off the shelf." Where an agency desires to engage the public with the design of a technical system, contracts[280] and

[276] Holdren & Smith, *supra* note 274, at 2 & n.2 (citing President Obama's remarks on November 23, 2009 at the launch of his "Educate to Innovate" campaign for excellence in science, technology, engineering, and math education).

[277] *See Big Data: A Tool for Inclusion Or Exclusion?*, FED. TRADE COMMISSION (Sept. 15, 2014), https://ftc.gov/news-events/events-calendar/2014/09/big-data-tool-inclusion-or-exclusion [https://perma.cc/69PE-4ZGH]; *Data Brokers: A Call for Transparency and Accountability*, FED. TRADE COMMISSION (May 2014), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf [https://perma.cc/6QSU-SD8A].

[278] *See The National Artificial Intelligence Research and Development Strategic Plan*, NAT'L SCI. & TECH. COUNCIL, at 8–40 (Oct. 2016), https://nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf [https://perma.cc/JBV5-Z5HK].

[279] *See, e.g.*, NAT'L ACADS. SCIS., ENG'G & MED., PROACTIVE POLICING: EFFECTS ON CRIME AND COMMUNITIES (David Weisburd & Malay K Majmundar eds., 2018), https://doi.org/10.17226/24928 [https://perma.cc/7A3H-USQV] (discussing biases in statistical predictions).

[280] *See* Joseph Lorenzo Hall, *Contractual Barriers to Transparency in Electronic Voting*, PROC. 2007 USENIX/ACCURATE ELECTRONIC VOTING TECH. WORKSHOP (2007), https://www.usenix.org/legacy/event/evt07/tech/full_papers/hall/hall_html/jhall_evt07_

intellectual property[281] may present formidable obstacles. Testing—a tried and true method of exploring how a system performs for various populations or under different conditions (usability) as well as security properties—may be contractually limited.[282] Companies may even demand secrecy of training materials. While revealing source code may not be a necessary or desirable means of exposing the values embedded in systems, allowing regulators and experts of their choosing to examine, test, and tinker with systems is an initial prerequisite for surfacing values for public consideration.

As noted elsewhere, public participation is important during both design and deployment.[283] The U.S. Public Participation Playbook provides best practices for agencies to engage the public. While the strategies it has identified through collaboration with agency staff and citizen engagement experts are technology agnostic, they provide guidelines for and examples of successful citizen engagement. Of particular importance for machine learning is the Playbook's mindset of ongoing public engagement and feedback, its emphasis on identifying engagement strategies for specific stakeholder groups, and its emphasis on using prototypes and systems as well as more traditional means of eliciting feedback.

Combining the high-level airing of issues in the White House reports discussed above with participatory processes at the USDS and 18F could provide a powerful first step in a framework for public engagement with the politics of machine learning systems. The development of the Web Standards by USDS and 18F was done through an open process and produced publicly available resources for both use and interrogation. In the words of USDS and 18F, "we're working in the open to create a resource that everyone can own and contribute to."[284] Admittedly, this sort of openness is not conducive to broad participation by the lay public, but it does open up technical design for interrogation by experts who can support such participation and deliberation.

However, USDS and 18F go beyond such technocratic openness. For example, 18F played a key role in implementing the Digital Accountability

html.html [https://perma.cc/GUY2-HY4X] (discussing the use of contracts to limit public oversight).

[281] *See, e.g.*, Aaron Burstein et al., LEGAL ISSUES FACING ELECTION OFFICIALS IN AN ELECTRONIC-VOTING WORLD (Mar. 15, 2007), https://www.law.berkeley.edu/files/Legal_Issues_Facing_Election_Officials.pdf [https://perma.cc/7DKX-JPG8]; Levine, *supra* note 114.

[282] Hall, *supra* note 280.

[283] *See* Mulligan & Bamberger, *supra* note 12, at 772 (arguing that "[t]he traditional sequential perspective of 'policymaking' (during which there is an opportunity for input) followed by 'implementation' is inconsistent with [regulation through] design").

[284] Ruskin, *supra* note 238.

and Transparency Act,[285] building a prototype of the proposed technical implementation to facilitate public feedback. In effect, through prototyping, 18F brought public participation into the agile software design process.[286] The prototyping used in this instance was one of many strategies for fostering and improving public participation in shaping government programs, including regulations, developed by 18F in conjunction with numerous government agencies.

Yet the broader public must be brought in, as well. The Federal Trade Commission's practices used to develop policy around the privacy impact of technology use and design model suggest two important elements of agency process: publicity and engagement.[287] The Commission conducted and publicized research on both child-directed and general audience websites, increasing the transparency of technology privacy behavior and its policy implications, spurring public attention and providing the basis for engagement. The participatory fora that followed empowered privacy advocates (indeed, many advocacy organizations were founded in response to this opportunity) who both provided input and served as a means for publicizing policy threats more broadly.

Meaningful public participation focused more directly on the adoption of specific systems requires similar scaffolding by agencies. Prototypes and simulations can be powerful means of publicizing technology choices by translating between mathematical formulations and policy choices. For example, the debate about how to model fairness in COMPAS (described above) may seem esoteric to members of the public, but a simulation that allowed individuals to play with various fairness metrics could powerfully expose them to the political implications. An example of such a tool is Google's What-if Tool. The What-if Tool is an interactive visual interface that allows individuals to probe machine learning models. The tool provides options to explore various algorithmic fairness constraints, compare counterfactuals, selectively add and remove data points, compare models on a single data set, among others.[288] Interactive exploratory tools provide opportunities for experiential learning by the public, building greater understanding of aspects of model design, and potentially framing questions about specific model design or data selection choices facing an agency.

[285] Digital Accountability and Transparency Act of 2014, Pub. L. No. 113-101, 128 Stat. 1146.

[286] *See Implementing A Government-wide Law*, 18F, https://18f.gsa.gov/what-we-deliver/data-act/ [https://perma.cc/WAC6-7FKF] (last visited Oct. 9, 2019).

[287] PRIVACY ON THE GROUND, *supra* note 228, at 189–90 (discussing FTC efforts).

[288] WHAT-IF TOOL, https://pair-code.github.io/what-if-tool/ [https://perma.cc/HVJ6-HHWN] (last visited Oct. 9, 2019).

The models of the Oakland Privacy Advisory Committee, the New York City Algorithmic Taskforce, and the Pennsylvania Sentencing Commission, moreover, use public events in which experts engage with impact assessments and systems before the public to further elaborate the policies that warrant attention. In effect, these models bolster the technological expertise of the public. Given that civil society organizations often have limited technical capacity, government provisioning of such experts to the public and stakeholders through expert committees may be an important component of the public participation infrastructure. Without agency enlistment of such expertise for the public, the public may miss risks and opportunities posed by machine learning systems and be unable to formulate appropriate and viable solutions.

3. *Contestable Design*

To ensure that the informed agency engages with and deliberates about the policymaking that occurs through system deployment and use, government machine learning systems must be designed to promote contestability.[289] That is, they must be designed to reveal their "thinking" and receive feedback from and collaborate with human users at runtime.[290] By fostering user engagement within the system, contestable systems use that engagement to iteratively identify and embed domain knowledge and contextual values as decision making becomes a collaborative effort in sociotechnical systems. Contestable systems thus provide a means for system logic to be overseen by, to learn from, and to be shaped by, the domain

[289] *See* Tad Hirsch et al., *Designing Contestability: Interaction Design, Machine Learning, and Mental Health*, 2017 PROC. ACM DESIGNING INTERACTIVE SYSS. CONF. 95, 98 (2017) (identifying "contestability" as a new design concern—focused on anticipating and designing for the ways technology can reshape knowledge production and power and describing three lower-level design principles to support contestability: 1) improving accuracy through phased and iterative deployment with expert users in environments that encourage feedback; 2) heightening legibility through mechanisms that "unpack aggregate measures" and "trac[e] system predictions all the way down" so that "users can follow, and if necessary, contest the reasoning behind each prediction"; and relatedly, in an effort to identify and vigilantly prevent system misuse and implicit bias, 3) identifying "aggregate effects" that may imperil vulnerable users through mechanisms that allow "users to ask questions and record disagreements with system behavior" and engage the system in self-monitoring).

[290] Contestable design aligns with Mireille Hildebrandt's call for " 'agonistic machine learning,' i.e., demanding that companies or governments that base decisions on machine learning must explore and enable alternative ways of datafying and modelling the same event, person or action." Mireille Hildebrandt, *Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning*, 20 THEORETICAL INQUIRIES L. 83, 106 (2019). This "highlights the contestability at the level of the inner workings of machine learning systems." *Id.* at 118. Also, this "responds to the need to call out the ethical and political implications of *who decides task T, performance metric P and experience E, and to investigate how this is done, taking into account which (and whose) concerns are at stake.*" *Id.* at 110.

expertise and experience of human agency staff actually vested with administrative discretion. Contestability is thus one way "to enable responsibility in knowing," to use Judith Simon's phrase, as the production of knowledge is spread across humans and machines.[291]

As legal scholars, we are reluctant to argue for generalized design choices for agency machine learning systems. However, existing literature suggests that particular design choices support contestability by fostering user understanding of models and outputs, collaborative construction of systems, dynamic feedback and control, and within-model challenges to system outputs.[292] These approaches suggest the beginnings of a list of design requirements that permit contestability, which frames an agenda for research and experimentation about governmental systems looking forward.

[291] Judith Simon, *Distributed Epistemic Responsibility in a Hyperconnected Era*, *in* THE ONLIFE MANIFESTO: BEING HUMAN IN A HYPERCONNECTED ERA 145–59, 146 (Luciano Floridi ed., 2015), https://doi.org/10.1007/978-3-319-04093-6_17 [https://perma.cc/2BQV-DBEG].

[292] For insights on how contestable systems advance individual understanding, see, for example, Motahhare Eslami, *Understanding and Designing Around Users' Interaction with Hidden Algorithms in Sociotechnical Systems*, 2017 CSCW '17 COMPANION 57 (describing several studies finding that seamful designs, which expose algorithmic reasoning to users, facilitated understanding, improved user engagement, and in some instances altered user behavior); Motahhare Eslami et al., *"I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds*, 2015 PROC. 33RD ANN. ACM CONF. ON HUM. FACTORS COMPUTING SYSS. 153 (describing the lasting effects on how users engage with Facebook to influence the News Feed algorithm after an experimental design intervention that visualized its curatorial voice); Susan Joslyn & Jared LeClerc, *Decisions with Uncertainty: The Glass Half Full*, 22 CURRENT DIRECTIONS PSYCHOL. SCI. 308 (2013) (describing how displaying uncertainty in weather predictions can lead to more optimal decision making and trust in a forecast: transparency about probabilistic nature of prediction engenders trust even when predictions are wrong); Simone Stumpf et al., *Toward Harnessing User Feedback For Machine Learning*, 2007 PROC. 12TH INT'L CONF. ON INTELLIGENT USER INTERFACES 82; Simone Stumpf et al., *Interacting Meaningfully with Machine-Learning Systems: Three Experiments*, 67 INT'L J. HUMAN-COMPUTER STUD. 639 (2009) (noting that explainable systems can improve user understanding and use of system and enable users to provide deep and useful feedback to improve algorithms); Travis Moor et al., *End-User Debugging of Machine-Learned Programs: Toward Principles for Baring the Logic*, SEMANTIC SCHOLAR (2009), https://pdfs.semanticscholar.org/9a4f/a9f116668927575113d6d2e8572e39925650.pdf?_ga=2.190583149.793734117.1566762743-966572889.1566762743&_gac=1.216591076.1566762743.CjwKCAjw44jrBRAHEiwAZ9igKEyD8gmFQSr6pK9M8nWVxjHzez0odTFCvRLu4dS7rXCjRvZGxRwsmBoCVXwQAvD_BwE [https://perma.cc/E7BA-MXS3] (noting that salient explanations helped users adjust their mental models); Saleema Amershi et al., *Power to The People: The Role of Humans in Interactive Machine Learning*, 35 AI MAG. 105 (2014) (providing an overview of interactive machine learning research with case studies, and discussing value of interactive machine learning approaches for the machine learning community as well as users).

a) Design Should Expose Built-in Values

At its core, contestability requires systems to be designed in a way that exposes value-laden features and parameters. Most simply, this visibility can prompt awareness, reflection, and feedback by agency decision makers relying on these systems. Reminding agency staff about the original choice of model and training data can prompt future deliberation about their appropriateness, particularly where a system consistently produces outcomes that agency staff perceive as incorrect in certain classes of cases. Decision makers relying on recidivism risk systems, such as Northpointe's COMPAS, must be made fully aware of how gender and other protected attributes are used, to both educate them about the policy choices underlying their decision supports and to encourage feedback from ground-level staff about their strengths and limitations. Brauneis and Goodman underscore the ways that contestability can foster ongoing deliberation about policy, describing the ways that the Hunchlab predictive policing software

> allows each community to set weights for the relative seriousness of each type of crime—how much more important is it to stop a murder than a burglary? It also allows tailored weights for patrol efficacy—[for example,] indoor crimes are less likely to be deterred by increased police presence.[293]

Moreover, design decision visibility can enable active participation by system users in consequential decisions about their configuration or use. For example, the confidence thresholds that determine an agency's preference for false positives or false negatives when using Amazon's Rekognition Web Service should be prominently exposed to staff and easily configurable. In these ways, contestable design expands front-line staff's knowledge of the policies and values embedded in the machine learning systems they use, while offering them opportunities to configure and interrogate them at run time. Such designs help agency staff learn about machine learning systems as the systems learn about agency staff. Machine learning systems, then, should be designed to allow users to both make key decisions about values-significant parameters and understand their significance. This requires moving away from defaults for these parameters and toward contestable systems that require engagement during setup and use.

Contestable design, moreover, is a prerequisite for continuous feedback from domain experts. Rather than traditional forms of contesting automated decisions—"out-of-band" processes (those external to the regular operation of the system itself, like exception handling and appeals)—contestable design brings argumentation, and therefore opportunities for learning and

[293] Brauneis & Goodman, *supra* note 6, at 150.

recalibration within the system itself. Such continuous within-system learning is appropriate as "our models are, and will continue to be, fallible" and is particularly important in areas where the risks of " 'getting it wrong' can be quite high."[294] Active, critical, real-time engagement with the reasoning of machine learning systems' inputs, outputs, and models reduces the risk that machine learning systems will replace the logical and ethical frameworks that comprise expert agency judgment, and respond to risks posed by fallible models—regardless of whether fallibility is a product of design choices, shifts in policy, or inattention to gaps between phenomena and the representations we choose to capture them.

Contestability is a more active and dynamic principle than explanation[295] to guide design; for these reasons it breeds user engagement. Where the passivity of "explainable" algorithmic systems imagines engagement, reflection, and questioning as out-of-band activities—via exception handling, appeals processes, etc.—contestable systems foster active, critical engagement within the system. Explanations are also typically static, insofar as they are focused on conveying a single message; contestability, in contrast, aims to support interactive exploration of, and in some instances tinkering with, machine logic.

b) Design Should Trigger Human Engagement

Designing for contestability further requires the design of human-system interaction in a way that promotes an active and ongoing role for agency decision makers by overcoming the over-reliance on, and deference to, decision-support systems arising from automation bias—"the use of automation as a heuristic replacement for vigilant information seeking and processing"[296] and automation complacency—insufficient attention and monitoring of automation outputs.

It is certain that without thoughtful interventions, agency staff will be less attentive and engaged with the decisions supported by machine learning systems. Where the goal of introducing machine learning systems is to achieve a hybrid production of knowledge that builds on the strengths of human and machine ways of knowing—rather than to fully displace human decision making—chosen designs and policies must keep humans "in the game."

[294] Hirsch et al., *supra* note 289, at 97.

[295] *See id.* at 98; *see also* Daniel Kluttz et al., *Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions*, *in* AFTER THE DIGITAL TORNADO: NETWORKS, ALGORITHMS, HUMANITY (Kevin Werbach ed., forthcoming 2020).

[296] Kathleen L. Mosier & Linda J. Skitka, *Automation Use and Automation Bias*, PROC. HUM. FACTORS & ERGONOMICS SOC'Y ANN. MEETING 344 (1999).

Research has established that particular design choices determine the extent to which expert judgment continues to be exercised in system-supported decision making in ways that skew policy regarding, for example, a preference for false-negatives over false-positives. The extent to which users substantively review a machine learning system output depends largely on whether the system signals that the result is anomalous or normal. For example, research has shown that radiologists scrutinize mammography films identified by decision-support systems as positive for cancer, catching many of the false positives that the system produced.[297] But films identified as normal rarely receive such inspection, allowing nearly all false negatives to evade detection. The perception that the system was over-inclusive—supported by the experience of identifying false positives—contributed to a belief in the infrequency of false negatives, when in fact the system systematically failed to identify certain images of cancer.

Similarly, in ongoing research with legal professionals using machine learning systems to aid in document review for discovery,[298] we have found that human review focuses on documents identified as relevant and thus appropriate for disclosure, according less attention to those the machine learning system designates as non-responsive. In both instances, perceptions of the machine's performance interact with experts' risk models to yield different levels of engagement with different outputs of the same underlying system. Yet, ideally we would want agency experts to be attentive to misidentifications or classifications (depending upon the task) produced by false positives and negatives—although, depending upon the use of the system, attention to one or the other might be directed by policy.

Iterative re-delegation of tasks and communicating about uncertainty are both design strategies that have been found successful as to maintain human skill and sense of responsibility and maintain attention to how machines are executing delegated tasks. For example, adaptive allocation,[299] in which an automated task is reallocated back to a human for periods of time, has been found to reduce automation complacency and improve subsequent attention

[297] Anrey A. Povyakalo et al., *How to Discriminate Between Computer-Aided and Computer-Hindered Decisions: A Case Study in Mammography*, 33 MED. DECISION MAKING 98 (2013); *see* Goddard, *supra* note 182; *see also* Adrian Bussone et al., *The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems*, INT'L CONF. ON HEALTHCARE INFORMATICS 160 (2015) (discussing research findings on automation bias and self-reliance).

[298] Daniel N. Kluttz & Deirdre K. Mulligan, *Automated Decision Support Technologies and the Legal Profession*, 34 BERKELEY TECH. L.J. 853 (2019).

[299] *See* RAJA PARASURAMAN ET AL., THEORY AND DESIGN OF ADAPTIVE AUTOMATION IN AVIATION SYSTEMS, REP. NO. NAWCADWAR-92033-60 (July 17, 1992), https://apps.dtic.mil/dtic/tr/fulltext/u2/a254595.pdf [https://perma.cc/YBJ6-UAXQ].

to tasks by humans,[300] reduce human distraction, and promote a sense of responsibility.[301] Communicating the confidence that a system has in the conclusions it offered has been found to foster feelings of user responsibility[302] and, where coupled with a feedback loop, improve decisions in the moment and over time.[303]

c) Design Should Promote Contestation About Social and Political Values

Finally, for a number of reasons, contestability is of heightened importance where functions delegated to machine learning systems are deeply connected to social and political values such as fairness—values that are often ambiguous and contested.

First, given the numerous competing definitions of fairness,[304] there may well be multiple conflicting views on which definition should apply to a given function or in a given context; and when a definition of fairness is chosen, it may be susceptible to different formalizations.[305] We have more generally

[300] *See* Raja Parasuraman, *Effects of Adaptive Task Allocation on Monitoring of Automated Systems*, 38 HUM. FACTORS 665 (1996).

[301] *Id.*

[302] Matthew Kay et al., *When (ish) Is My Bus?: User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems*, PROC. 2016 CONF. ON HUM. FACTORS COMPUTING SYSS. 5092 (2016) (finding that some bus riders felt that providing information about the uncertainty of an arrival time would make them more responsible for actions that led them to miss the bus: "you're more likely to be unhappy than if you missed the bus and can just blame the app").

[303] Communicating the uncertainty related with the probabilistic nature of machine learning systems to users improves decision making, and if it is coupled with feedback mechanism, it can leverage human knowledge to increase accuracy of model over time. *See* Rocio Garcia-Retamero & Edward T. Cokely, *Communicating Health Risks with Visual Aids*, 22 CURRENT DIRECTIONS IN PSYCHOL. SCI. 392–93 (2013) (displaying a grid of pictograms, each representing a patient success or fatality improved the accuracy of people's risk assessment); *see also* Ulrich Hoffrage & Gerd Gigerenzer, *Using Natural Frequencies to Improve Diagnostic Inferences*, 73 ACAD. MED. 538 (1998) (noting that more medical experts could accurately estimate the positive predictive value (precision) of a test when presented with discrete counts or outcomes); Stumpf et al., *Toward Harnessing*, *supra* note 292, at 82 (noting that "user feedback has the potential to significantly improve machine learning systems").

[304] *See* Deirdre K. Mulligan et al., *This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology*, PROC. 2019 ACM ON HUMAN-COMPUTER INTERACTION 3, 119 (2019) (defining the following terms: formal equality (blind to all other variables)—to each person an equal share; need-based—to each person according to individual need; effort-based—to each person according to individual effort; social contribution—to each person according to societal contribution; and merit-based—to each person according to merit).

[305] For example, one could operationalize a given fairness definition around groups—seeing demographic parity or equal positive predictive values, or equal negative predictive values, or equal false positive or false negative rates, or accuracy equity—or one could

argued that, because of the strength and durability of design decisions, policymakers should be cautious about "baking" choices about human and public rights into technology systems and should steer such determinations to the least fixed point of technical intervention—permitting ongoing debate about such value choices and "designing technological hooks that permit different value choices in different contexts."[306] The discussion of the debates around what fairness required in COMPAS offers a clear example of the need to maintain visibility about these contested design elements.

Second, protecting and respecting values such as fairness and privacy may often hinge on *process*. Fairness and privacy, for example, are often *defined*—at least in part—by access to procedures that afford individuals meaningful participation, and to information (rules and data used to make decisions about them). These procedural aspects of values can be supported through contestable system design that minimizes the automation and opacity of those decisions and ensures that human judgment will be brought to bear in the ongoing shaping of, and in the assessment of the products of, decision-support systems.[307] Contestability keeps agency experts in control of these values questions, even while specific tasks and functions are handed off to machine learning systems. They can allow agency experts to revisit and tune machine-learning decisions to context-specific information that influences the perceived or actual fairness of a system. Contestable design responds to the demand of philosophers and ethicists that systems be designed to respond to diverse contexts ruled by different moral frameworks[308] and to support collaborative development of ethical requirements.[309]

V. CONCLUSION

In 1967 John Culkin, interpreting one of Marshal McLuhan's five postulates, offered the now-famous line: "We shape our tools and thereafter

operationalize it through an individual fairness metric, such as equal thresholds or devising a similarity metric. For a discussion, see *id.*

[306] Mulligan & Bamberger, *supra* note 12, at 750.

[307] *See* Min Kyung Lee & Su Baykal, *Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division*, PROC. 2017 ACM CONF. ON COMPUTER SUPPORTED COOPERATIVE WORK & SOC. COMPUTING 1035 (2017) (presenting discussion with others, and ability to interrogate systems logic can improve perceptions of fairness).

[308] *See* Batya Friedman & Helen Nissenbaum, *Software Agents and User Autonomy*, PROC. 1ST INT'L CONF. ON AUTONOMOUS AGENTS 466 (1997).

[309] *See* Matteo Turilli, *Ethical Protocols Design*, 9 ETHICS & INFO. TECH. 49 (2007).

they shape us."[310] The fear of abdicating important policy decisions to the control of tools, rather than autonomously wielding them in service of the public interest, lies at the heart of current concerns with the governmental adoption of machine learning systems. Such abdication strikes at the heart of administrative legitimacy and good governance, and suggests that machine learning systems are tools to be procured at our peril.

In many instances they are more texts than tools, and we suggest engaging them accordingly. They have their limits, their biases, and their blind spots. We should question and bicker with them, but we should also learn from and teach them. We should not blindly defer to them or bring them into our deliberations without knowing their backstories, working assumptions, and theories.

Through policy choices and design, we can build purposeful tools that are aligned with values chosen based on reason, expertise, transparency, and robust and ongoing deliberation and oversight. We can bind these new systems through carefully constructed "policy knots" that align them with the requirements of administrative law through policy, practice, and design.

[310] John M. Culkin, *A Schoolman's Guide to Marshall McLuhan*, SATURDAY REV. 51, 70–72 (Mar. 18, 1967) (offering a "barously brief" distillation of Marshal McLuhan's writings, John M. Culkin expanded on one of McLuhan's five postulates, *Art Imitates Life*).