Harvard Journal of Law & Technology
Volume 37, Number 1 Fall 2023

ALGORITHMIC RULEMAKING VS. ALGORITHMIC GUIDANCE

Peter Henderson & Mark Krass***

ABSTRACT

Algorithms are coming to government. One legal question raised by this change is the extent to which the Administrative Procedure Act ("APA") will regulate the use of algorithms as decision-support tools for agency adjudicators. Under the APA, "rules" are officially binding statements of policy subject to notice and comment as well as rigorous pre-implementation judicial review, whereas "guidance," officially defined as non-binding advice, is effectively unreviewable. The implementation of algorithmic tools often occupies a gray zone between the two. To help clear the thicket, we provide a deep dive into the computer science and economics literature to provide a set of workable heuristics that help distinguish algorithmic rulemaking from algorithmic guidance. These heuristics align with best practices in the computer science literature and provide insights into agency incentives for adopting safer algorithms. We suggest that the specter of rulemaking may have value in nudging agencies toward best practices aligned with existing algorithmic safety recommendations. Specifically, avoidance of APA rulemaking may encourage agencies to prevent automation bias and other potential harms from algorithmic deployments. In this way, distinguishing algorithmic rules and guidance under the existing framework of the APA may dovetail with best practices in computer science.

* Assistant Professor in the Department of Computer Science and School of Public and International Affairs at Princeton University.

** Stanford Department of Political Science, Ph.D. Candidate. This paper reflects our personal views only and not those of any company, government entity, institution, or other person. We are grateful to Ryan Calo, Amit Haim, Daniel Ho, David Hausman, and participants in WeRobot 2022 for their helpful suggestions. We also thank the staff of the *Harvard Journal of Law & Technology* for their excellent editorial assistance. The authors contributed equally and are listed in alphabetical order; all errors are our own.

TABLE OF CONTENTS

I. INTRODUCTION

Algorithms are coming to government. Many scholars have offered thoughtful proposals on how to regulate that sea change.[1] But those proposals may take years to bear fruit. This Article asks how courts, agencies, and regulated parties might use an existing tool, the

[1] *See, e.g.*, DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO-FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES (2020), https://law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf [https://perma.cc/ESJ4-UWWF]; CARY COGLIANESE, A FRAMEWORK FOR GOVERNMENTAL USE OF MACHINE LEARNING 62–75 (2020), https://www.acus.gov/sites/default/files/documents/Coglianese%20ACUS%20Final%20Report%20w%20Cover%20Page.pdf [https://perma.cc/BFT7-3YZQ].

Administrative Procedure Act ("APA"), to regulate algorithms in government.[2]

Here's the basic issue: In some critical applications, algorithms do not produce binding decisions; rather, they are used to support human decision-makers by giving them recommendations. This places algorithms, and the documents requiring agency officials to consult their outputs, in a doctrinal gray zone between *rules* and *guidance*. The dividing line between these categories can be summed up in one word: "discretion."

Under Section 553 of the APA,[3] "rules" (sometimes called "legislative rules") remove discretion from agency adjudicators; as a result, rules are subject to both notice and comment and pre-implementation review by courts.[4] On the other hand, "guidance" (in the language of Section 553, a "general statement of policy") is take-it-or-leave-it advice that does not bind agency officials when making decisions and is effectively unreviewable under the APA, although other regulatory schemes might still apply.[5] By drawing on existing economics and computer science literature on the integration of informational signals into decision-making, this Article offers a new resource for defining what discretion looks like in the context of algorithmic support tools.

The meaning of discretion, and by extension the APA's distinction between rules and guidance, will help shape the regulation of algorithms in government. Treating algorithms as "rules" and subjecting them to notice and comment would put plaintiffs and courts in the driver's seat, allowing them to demand extensive factual development,[6]

[2] Our focus on extant tools requires us to set aside a few important categories of algorithms. The APA, and separation-of-powers principles, have made courts wary of supervising most enforcement decisions, so we leave those to the side for purposes of this piece, despite our misgivings about the wisdom and legal underpinnings of that reticence. *See* United States v. Texas, 143 S. Ct. 1964, 1971 (2023) ("Under Article II, the Executive Branch possesses authority to decide how to prioritize and how aggressively to pursue legal actions against defendants who violate the law.") (internal quotation omitted); Heckler v. Chaney, 470 U.S. 821, 831 (1985) ("This Court has recognized on several occasions over many years that an agency's decision not to prosecute or enforce, whether through civil or criminal process, is a decision generally committed to an agency's absolute discretion."); David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 YALE J. ON REG. 800, 830 (2020). Likewise, we set aside algorithms targeted at strictly internal activities like research or management because these, too, are subject to an APA exception, *see* 5 U.S.C. § 551 (defining an "agency action"), and because no plaintiff is likely to have standing to sue.

[3] 5 U.S.C. § 553.

[4] *See* Nicholas R. Parrillo, *Federal Agency Guidance and The Power to Bind: An Empirical Study of Agencies & Industries*, 36 YALE J. ON REG. 165, 167–68 (2019).

[5] *Id.*

[6] APA challenges require agencies to produce the complete record reflecting all of the information they used to make a given decision. *See, e.g.*, Motor Vehicles Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 43–44 (1983) ("Congress required a record of the rulemaking proceedings to be compiled and submitted to a reviewing court[.]"); Aram A. Gavoor & Steven A. Platt, *Administrative Records After* Dep't of Commerce v. New York, 72 ADMIN. L. REV. 87, 98 (2020).

imposing the APA's basic preference for stability over change,[7] and potentially facilitating non-APA suits through the discovery process.[8] But if all algorithms are mere "guidance," executive oversight will take the primary role in regulation and the public's access to information will instead rely on uncertain mechanisms such as Freedom of Information Act ("FOIA") litigation, agency inspectors general, and executive grace.[9] Whatever one might think of granting the executive branch autonomy to craft its own transparency policy on algorithms, initial indications on the efficacy of that approach are not encouraging.[10] So while we take on board critiques of the APA as a framework for regulating artificial intelligence ("AI") in government, we think it quite

[7] *See* Engstrom & Ho, *supra* note 2, at 839 ("Notice and comment is a protracted process and, when combined with pre-enforcement review, can stymie innovation and prevent dynamic government responses to a changing policy problem or regulatory landscape.").

[8] For example, the availability of APA suits might bring to light the evidence needed to establish standing or survive a motion to dismiss in an antidiscrimination or privacy suit. In the absence of such evidence, plaintiffs might face the same catch-22 as those litigating over the National Security Agency's wiretapping program, where courts denied discovery because plaintiffs had no seed evidence to bolster their complaints. *See* Clapper v. Amnesty Int'l USA, 568 U.S. 398, 410–12 (2013).

[9] The Freedom of Information Act, 5 U.S.C. § 552, will sometimes provide an alternative route to getting information about algorithms. *See* Nat'l Res. Def. Council v. EPA, 954 F.3d 150, 152 (2d Cir. 2020) (requiring the EPA to release a model predicting greenhouse gas emissions under FOIA). True, FOIA's deliberative process exemption, which shields from disclosure documents that form "part of a process by which governmental decisions and policies are formulated," will also therefore protect the recommendations of the most guidance-like algorithms. Elec. Frontier Found. v. DOJ, 739 F.3d 1, 7 (D.C. Cir. 2014) ("The privilege is limited to documents that are 'predecisional' and 'deliberative,' meaning 'they reflect[] . . . recommendations . . . comprising part of a process by which governmental decisions and policies are formulated.'") (internal citations omitted). On its face, though, the deliberative process exemption would not shield information about the models underlying those recommendations from disclosure, since only model outputs are predecisional in the relevant sense. Doubly so, given that Congress has recently required agencies to compile and make public "an inventory of the artificial intelligence use cases" they have deployed. James M. Inhofe National Defense Authorization Act for Fiscal Year 2023, Pub. L. No. 117-263, § 7225(a)(1)–(3), 136 Stat. 2395, 3671–72 (2022). Still, FOIA litigation is tilted against oversight and may prove inadequate given the political and commercial sensitivity of new algorithmic tools. *See generally* Margaret B. Kwoka, *FOIA, Inc.*, 65 DUKE L.J. 1361 (2016) (discussing the domination of FOIA by corporate requesters). Likewise, agencies' internal inspectors general are powerful in principle but are often stymied by agency resistance (and of course restricted by the inspector general's own competence and willingness to investigate). *See* Shirin Sinnar, *Protecting Rights from Within: Inspectors General and National Security Oversight*, 65 STAN. L. REV. 1027, 1031–32 (2013).

[10] *See* CHRISTIE LAWRENCE, ISAAC CUI & DANIEL E. HO, IMPLEMENTATION CHALLENGES TO THREE PILLARS OF AMERICA'S AI STRATEGY 13 tbl.1 (2022), https://dho.stanford.edu/wp-content/uploads/AI_Implementation.pdf [https://perma.cc/8MWR-QKWJ] (reporting that seventy-six percent of parent and sub-agencies potentially subject to Exec. Order No. 13,960, 85 Fed. Reg. 78939 (2020), when assessed separately, have failed to comply with the Order's requirement that they publish an inventory of implemented algorithms).

plausible that the APA will, in fact, matter a great deal.[11] Here, we focus our efforts on understanding how the APA's doctrinal categories might be adapted to the use of algorithms in government.

Observing discretion is difficult. Legions of commentators have bemoaned the "remarkably tough" task of distinguishing rules and guidance.[12] It is easy to see why. Imagine that the Department of Justice ("DOJ") contracts to build a decision-support tool for immigration judges ("IJs") that suggests whether or not to grant asylum. The agency issues a short memo "recommending" that IJs consult the tool's outputs. Assume we know that IJs defer to the tool in one hundred percent of cases, even though they theoretically retain discretion to weigh the evidence as they see fit. Was DOJ's recommendation letter a "rule"?

Formalists would have an easy answer: No. As long as immigration judges retain the power to reject an algorithm's recommendation on paper, they are not subject to a rule. But the APA's text seems to foreclose such a facile response. It defines any statement "*designed* to implement, interpret, or prescribe law or policy" as a rule, phrasing that implicitly points to the real-world effects of an agency policy.[13] Worse, a formalist test would give agencies an easy cheat code to avoid costly APA review.

Unsurprisingly, the D.C. Circuit has expressly rejected that approach, opting instead to look at the function of an agency document in practice to decide whether it is a rule or guidance document. Specifically, in *Appalachian Power Co. v. EPA*,[14] the court held that if an agency "acts as if a document issued at headquarters is controlling in the field"; "bases enforcement actions on the policies or interpretations" in the document; and leads citizens to "believe" that it plans to enforce the document as written, then it is "for all practical purposes 'binding.'"[15] Here, we follow the D.C. Circuit's approach and assume that a court will look at the agency's practices in the real world to decide whether a document is a rule or guidance.

[11] *See* Engstrom & Ho, *supra* note 2, at 813, 836–40 for an extensive critique along these lines. Among the most persuasive reasons for worrying about the imposition of the APA is the possibility that it will lead agencies to avoid using algorithms at all — preferencing existing systems characterized by significant due process violations over potentially powerful new tools for improving agency performance. *See, e.g.*, David Ames, Cassandra Handan-Nader, Daniel E. Ho & David Marcus, *Due Process & Mass Adjudication: Crisis & Reform*, 72 STAN. L. REV. 1, 24 (2020) (discussing due process tradeoffs); Kurt Glaze, Daniel E. Ho, Gerald K. Ray & Christine Tsang, *Artificial Intelligence for Adjudication: The Social Security Administration & AI Guidance*, *in* OXFORD HANDBOOK OF AI GOVERNANCE (forthcoming Apr. 2024) (discussing the use of AI to improve accuracy in Social Security Administration hearings), https://dx.doi.org/10.2139/ssrn.3935950 [https://perma.cc/ZQ7A-CL9H].

[12] Parrillo, *supra* note 4, at 170–72 nn.20–21 (collecting sources); *see also infra* Section III.A; Ronald M. Levin, *Rulemaking and the Guidance Exception*, 70 ADMIN. L. REV. 263, 265–66 (2018).

[13] 5 U.S.C. § 551(4) (emphasis added).

[14] 208 F.3d 1015 (D.C. Cir. 2000).

[15] *Id.* at 1021.

It might be tempting to think that empirical data would be helpful in determining whether an agency document satisfies a functionalist test of the kind outlined in *Appalachian Power.* But that is not so clear. For one thing, APA challenges will often arise before implementation — when the only empirical data might come from settings far removed from the real-world, in-the-field context where an algorithm is deployed. The relevance and accuracy of that data might be hard to judge. The bigger problem is that without a theory of discretion, empirical data is often difficult to interpret. Recall our immigration example above, where IJs adopted the "recommendations" of the algorithm one hundred percent of the time. Does total deference to an algorithm demonstrate the wholesale abdication of adjudicatory discretion? Or does it reflect judges' rational agreement with, or reliance on, a perfectly calibrated system?

We help cut through this complexity in three ways. First, we propose a definition of discretion focused on an adjudicator's tendency to update a default option. When agency adjudicators retain genuine discretion, they expend effort to draw upon the factual record and their own judgment to change the default disposition of a case in some way, even if only to make the default outcome more certain to be correct. Conversely, adjudicators lack discretion when they do not affect the probability of a given outcome. If an algorithm is so perfect as to never need the human's involvement in adjudication, then the human has less discretion in the relevant sense.

Second, we draw from several areas of study to offer a set of heuristics that courts might draw upon to guess when an algorithmic tool is more likely to result in this kind of updating. We look principally to two fields: the economics literature on Bayesian persuasion and computer science literature on human-computer interaction. We offer an original view of how insights from these fields can help identify features of algorithms that are more likely to reduce adjudicator discretion, like the amount of effort required for contradicting an algorithmic recommendation and the degree of time pressure imposed on adjudicators. We also discuss how different algorithmic tools might interact with agency context, such as the costs of deviating from algorithmic recommendations and the relative cost of finding information.

Finally, we show how our approach to identifying decision-support regimes that reduce human involvement might align the incentives of agencies in helpful ways. For example, our observation is that algorithms are less likely to be practically binding if they take steps to prevent automation bias, such as pointing the adjudicator to sources of information instead of directly recommending an action. In this way, the specter of notice and comment rulemaking could push agencies toward adopting safer and more transparent algorithmic tools, in line with what experts in algorithmic safety recommend.

We emphasize again that our focus on the APA is not a suggestion that it is the optimal tool for regulating algorithms in government. Further, courts may want to consider factors other than discretion — like the significance of an algorithm's outputs for governing primary conduct — in deciding whether an algorithm should be subject to APA review.[16] For instance, an algorithm that autonomously scans and recognizes the addresses on incoming mail might leave little discretion in the context of a transcription decision, but it does not rise to the level of significance for the primary conduct of individuals that would necessitate notice and comment review. Taken in the broader context of adjudicatory decisions involving the address, it plays a minor informational role, leaving plenty of room for discretion.

But, for now at least, the APA with its principal focus on discretion is a fact of life.[17] Courts will have no choice but to grapple with the meaning of discretion in the context of governmental algorithms.[18] Cases like *Velesaca v. Decker*,[19] which examined the use of an adjudication algorithm by Immigration and Customs Enforcement ("ICE"), show that they already do.[20]

The Article proceeds as follows. In Part II, we start by reviewing the growth of algorithms in the administrative state and how previous scholars have approached the intersection of that transformation with administrative law. In Part III, we focus on our proposed definition of rules in the context of algorithms. In Part IV, we turn to identifying rules. We first introduce the literature on Bayesian persuasion and human-computer interaction, then draw upon principles from those fields to explain how particular algorithmic designs can affect adjudicator discretion. We then turn to external agency contextual factors that might interact with algorithmic design elements to induce greater reliance by line adjudicators. We conclude in Part V by emphasizing how the factors we highlight might provide a constructive impetus for agencies to adopt safer algorithms in order to avoid the strictures of notice and comment rulemaking. Finally, we summarize our proposed test for courts.

[16] *See* Engstrom & Ho, *supra* note 2, at 846.

[17] For a discussion of the thirteen-year fight to pass the APA, see Walter Gellhorn, *The Administrative Procedure Act: The Beginnings*, 72 VA. L. REV. 219 (1986). For a discussion of the profound challenges that have attended intervening efforts to pass across-the-board reform in the nearly eighty years since, see Stuart Shapiro & Deanna Moran, *The Checkered History of Regulatory Reform Since the APA*, 19 N.Y.U. J. LEGIS. & PUB. POL'Y 141 (2016).

[18] Of course, as we discuss below, the utility of the notice and comment procedure for introducing greater transparency to the adoption of algorithms is conditional on a sufficiently specific rulemaking. The few decision-support tools that are explicitly authorized by rules have been promulgated under such general rules that scrutiny of the particular algorithms was limited. *See* Engstrom & Ho, *supra* note 2, at 838. Nonetheless, ex ante review *may* provide courts with some leverage to demand evidence that helps characterize the algorithm's likely performance.

[19] 458 F. Supp. 3d 224 (S.D.N.Y. 2020).

[20] *Id.* at 239, 241–42.

II. The Algorithmic State and Administrative Law

The use of algorithms in government is expanding rapidly. Scholars have written a great deal about how the law ought to regulate these emerging tools. This Part summarizes both the factual trend and the prior work to which our paper responds. In brief, we add to the prior literature by adding depth to the project of rationally reconstructing the existing legal regime, rather than offering a proposal for something new.

A. The Rise of the Algorithmic State

Although the term "algorithm" can literally refer to any "step-by-step procedure,"[21] this paper is focused on the family of administrative procedures in which at least one decisional step is partially informed by machine learning or AI — that is, a computer-based model that maps data onto a recommendation about how to proceed.[22] A model's guess might be situated at many levels of generality. It might recommend relevant legal materials[23] or sift through an administrative record to identify particular documents that would be helpful to the adjudicator.[24] Or it might make a recommendation as to a final decision, such as whether to grant bail to a person accused of a crime[25] or remove a child from a

[21] *Algorithm*, Merriam-Webster, https://www.merriam-webster.com/dictionary/algorithm [https://perma.cc/NZ5H-GA72]. This would, of course, encompass all administrative procedures whether or not they depend on a model. *See, e.g.*, Anupam Chander, *The Racist Algorithm*, 115 Mich. L. Rev. 1023, 1031–32 (2017) (reviewing Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (2015)) (characterizing the Federal Sentencing Guidelines as "sentencing based on ranges determined by algorithm").

[22] This somewhat imprecise definition is drawn in part from Aziz Z. Huq & Mariano-Florentino Cuéllar, *Toward the Democratic Regulation of AI Systems: A Prolegomenon* 6 (Univ. Chi. Pub. L. Working Paper No. 753, 2020), https://dx.doi.org/10.2139/ssrn.3671011 [https://perma.cc/9BK8-J34C]. We note that this definition is inclusive of simple methods like decision trees and linear regressions, alongside more complicated and opaque systems like deep neural networks. Some have argued against using the term "artificial intelligence" for such approaches, so the rest of this paper adopts the term "algorithm" instead. *See, e.g.*, Kathy Pretz, *Stop Calling Everything AI, Machine-Learning Pioneer Says*, IEEE Spectrum (Mar. 31, 2021), https://spectrum.ieee.org/stop-calling-everything-ai-machinelearning-pioneer-says [https://perma.cc/A5PQ-P25Z].

[23] *See, e.g.*, Zihan Huang, Charles Low, Mengqiu Teng, Hongyi Zhang, Daniel E. Ho, Mark S. Krass et al., *Context-Aware Legal Citation Recommendation Using Deep Learning*, 2021 Proc. 18th ACM Int'l Conf. on AI & L. 79, https://dl.acm.org/doi/pdf/10.1145/3462757.3466066 [https://perma.cc/HJ73-CKXR].

[24] For a review of similar technology focused on the civil litigation context, see Neel Guha, Peter Henderson & Diego Zambrano, *Vulnerabilities in Discovery Tech*, 35 Harv. J.L. & Tech. 581 (2022).

[25] For a discussion of algorithms in criminal pretrial and parole proceedings, see Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig & Sendhil Mullainathan, *Human Decisions and Machine Predictions*, 133 Q.J. Econ. 237 (2018).

potentially abusive home.[26] In each of these cases, humans may combine the output of the model with other evidentiary inputs to reach a final decision on the matter at hand, although in the latter two cases merely relying on the output of the model might be sufficient to dispose of the case.[27] End-to-end automation would be the most extreme case of a model-based decision-support tool.[28]

A recent report prepared for the Administrative Conference of the United States documents the growth of these kinds of algorithmic systems in government and maps the typical settings in which they are situated.[29] As we note above, we focus our discussion on systems designed to aid in administrative adjudication, the better to avoid tricky legal questions associated with other domains like enforcement, research, or public communication.[30] To help animate the discussion that follows, we offer two examples of the kinds of algorithms we address in this paper:

(1) **ICE's** Risk Classification Assessment ("RCA") algorithm.[31] While line ICE officers are often involved in enforcement efforts, ICE also takes a first-line role in determining whether non-citizens whose removal proceedings are being

[26] *See, e.g.*, Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski & Shion Guha, *A Human-Centered Review of the Algorithms Used Within the U.S. Child Welfare System*, 2020 PROC. CONF. ON HUMAN FACTORS IN COMPUTING SYSTEMS, Apr. 2020, at 1, 1, https://arxiv.org/pdf/2003.03541.pdf [https://perma.cc/F75P-LH5X].

[27] Note that this excludes a number of machine learning tools that may be used to organize adjudicatory systems in ways invisible to individual adjudicators. For example, the Social Security Administration uses predicted case outcomes to determine the *order* in which cases are resolved, but the predictions are hidden from adjudicators. *See* ENGSTROM ET AL., *supra* note 1, at 40. This would be excluded from our scope. Since adjudicators do not perceive the output of the algorithm, it cannot form part of their prior belief set. For more detail, see Part III.

[28] To name just one difference, there may be little gap between "rulemaking" and "adjudication" in a system that is fully automated end-to-end, since a given set of facts might deterministically produce a specific result. *See* Danielle K. Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1253 (2008) ("Computer programs seamlessly combine rulemaking and individual adjudications without the critical procedural protections owed either of them."). This would produce doctrinal problems for administrative law that are beyond the scope of this paper to resolve. Though the focus of many papers, end-to-end automation appears very far from reality for the majority of applications where discretion is currently employed. For example, all of the examples listed in the Administrative Conference of the United States report rely on humans at some point before a final decision. *See* Engstrom & Ho, *supra* note 2, at 811–12.

[29] *See* ENGSTROM ET AL., *supra* note 1.

[30] We discuss the challenges of judicial review in the context of algorithms designed to assist with enforcement, research, and public communication in *supra* note 2. Of course, distinguishing enforcement from adjudication can be tricky. We think of enforcement as dealing with the prosecutorial function of initiating adversarial proceedings.

[31] This example is drawn from David Hausman, The Danger of Rigged Algorithms: Evidence from Immigration Detention Decisions (July 8, 2021) (unpublished manuscript), https://dx.doi.org/10.2139/ssrn.3877470 [https://perma.cc/7RNN-TJCV] (documenting the facts of *Velesaca v. Decker*, 458 F. Supp. 3d 224 (S.D.N.Y. 2020)).

adjudicated should be released on bail or detained.[32] Drawing from the data available in ICE's files, the RCA algorithm makes a recommendation as to whether the non-citizen should be released.[33] That decision is reviewed by a line officer, and the officer's decision is reviewed by a supervisor.[34]

(2) **Social Security Administration's** Insight System ("Insight").[35] The Social Security Administration's Office of Appellate Operations ("OAO") developed the Insight system to "flag potential policy compliance or internal consistency errors" in hearing decisions issued by administrative law judges ("ALJs").[36] Based on Insight's recommendation, among other factors, OAO's adjudicators could recommend that the agency "affirm, modify, reverse, or remand" ALJ hearing decisions.[37] The system was ultimately rolled out to line-level ALJs to flag issues before decisions were sent to OAO for subsequent review.[38] An example of the type of problem that Insight flags might be a citation to a nonexistent provision of the Code of Federal Regulations.[39]

B. How Should Government Algorithms Be Regulated?

Much of the work on algorithms in government is oriented toward policy. Many pieces focus on either (a) addressing the tension between algorithmic administration and constitutional values or (b) recommending governance regimes that would require statutory reform. By contrast, we focus on how courts would handle algorithms under the current doctrinal regime.

Two sets of constitutional values, due process and equal protection, take the spotlight in discussions of algorithms in government. The principal due process concern with AI is the opacity of models.[40] The lack

[32] *Id.*; *see also* Robert Koulish & Kate Evans, *Punishing with Impunity: The Legacy of Risk Classification Assessment in Immigration Detention*, 36 GEO. IMM. L.J. 1, 4 (2021) (describing the bail adjudication process and the role of the Risk Classification Assessment in this process).

[33] Koulish & Evans, *supra* note 32, at 4.

[34] *Id.*

[35] *See* Glaze et al., *supra* note 11, at 14.

[36] OFF. OF THE INSPECTOR GEN., SOC. SEC. ADMIN., A-12-18-50353, AUDIT REPORT: THE SOCIAL SECURITY ADMINISTRATION'S USE OF INSIGHT SOFTWARE TO IDENTIFY POTENTIAL ANOMALIES IN HEARING DECISIONS 1 (2019), https://oig-files.ssa.gov/audits/full/A-12-18-50353.pdf [https://perma.cc/6XQ4-UR6V].

[37] *Id.* at 2.

[38] *Id.* at 6.

[39] ENGSTROM ET AL., *supra* note 1, at 40.

[40] For an overview of this set of concerns, see Citron, *supra* note 28, at 1253–55 (summarizing concerns over accuracy, opacity, misdelegated power, and other due process issues

of public access to source code and other details about how models work internally is one due process issue.[41] Another is the paucity of reasoning available to decision-makers. Ryan Calo and Danielle Citron, for instance, argue that agencies' reliance on algorithms is an inherent threat to their legitimacy because it outsources expertise to the authors of the software; by the same token, judges do not have a meaningful decision to review if the true rationale of the decision is unknown.[42] We return to the theme of reason-giving in algorithms below, as we think that the notice and comment regime may push agencies toward less prescriptive and more interpretable model output. Models' accuracy is also a source of due process concern, as overreliance on flawed models may undercut the reasoned deliberation to which parties may be entitled.[43] Of course, to the extent that models are able to mimic an appropriately deliberative procedure, they may alleviate the due

related to computer-driven adjudicative processes). The taxonomy of opacity that follows is inspired by Cary Coglianese, *The Transparency President? The Obama Administration and Open Government*, 22 GOVERNANCE 529 (2009). *But cf.* Lilian Edwards & Michael Veale, *Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 67 (2017) (noting that certain types of transparency, in terms of algorithmic explanations, may not be the type of transparency that is desirable).

[41] A related proposal focuses on incorporating guarantees of "conform[ance] to specified standards for ethical AI" and "limited waivers of trade secret and other intellectual property protections" into public procurement efforts to ensure the transparency of algorithms acquired from contractors. Lavi M. Ben Dor & Cary Coglianese, *Procurement as AI Governance*, 2 IEEE TRANSACTIONS ON TECH. & SOC'Y 192, 195 (2021).

[42] *See* Ryan Calo & Danielle K. Citron, *The Automated Administrative State: A Crisis of Legitimacy*, 70 EMORY L.J. 797, 803–04 (2021); *see also* Andrew D. Selbst & Salon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1118–22 (2018) (asserting that algorithmic explanation is a "good unto itself" and "forces the basis of decision-making into the open"); Citron, *supra* note 28, at 1253 (describing how "lack [of] record-keeping audit trails" makes review of an automated system's decision-making "impossible"); Bernard W. Bell, *Replacing Bureaucrats with Automated Sorcerers?*, 150 DAEDALUS 89, 97 (2021) (discussing the need for rational reason-giving and its relationship to external review). *But see* Cary Coglianese, *Administrative Law in the Automated State*, 150 DAEDALUS 105, 108 (2021) ("An adequate explanation could involve merely describing the type of algorithm used, disclosing the objective the algorithm was established to meet, and showing how the algorithm processed a certain type of data to produce results that were shown to meet the algorithm's defined objective as well as or better than current processes."). Aziz Huq is a skeptic of the opacity critique, at least on normative grounds and possibly on legal grounds. *See* Aziz Z. Huq, *A Right to a Human Decision*, 106 VA. L. REV. 611, 643–46 (2020) (noting that human decisions are often also opaque in any relevant sense, so that the kinds of post-hoc decisions available for algorithmic decisions are similar to those given by humans to rationalize their choices).

[43] *See, e.g.*, Saar Alon-Barkat & Madalina Busuioc, *Human-AI Interactions in Public Sector Decision Making: 'Automation Bias' and 'Selective Adherence' to Algorithmic Advice*, 33 J. PUB. ADMIN. RSCH. & THEORY 153, 159–61 (2022) (documenting two experiments in which participants dramatically overestimated the accuracy of algorithmic predictions, especially when predictions matched priors). A related but distinguishable concern is that administrative agencies lack the capacity to properly supervise the implementation of algorithmic systems, though that worry sounds in many registers beyond accuracy alone. *See* Calo & Citron, *supra* note 42, at 845.

process concerns that exist at present in overburdened adjudicatory agencies.[44]

Discrimination on the basis of protected attributes is another source of concern. The effort to use algorithms to predict recidivism in the context of bail and parole hearings has demonstrated that real-world disparities in training data invariably translate into unfairness along some dimension when that data is used to train a model.[45] Expanding the challenge, efforts to remedy fairness concerns can have deleterious effects on other values like accuracy and privacy.[46] Training data may also be incomplete in ways likely to produce biased models.[47] While all of these concerns are present even outside of the government context, the concern that algorithmic tools could encode systemic biases is even sharper as applied to the government context because of the unique dignitary harms of officially sanctioned discrimination. As with due process concerns, the status quo — human adjudication — may be worse than the replacement, but the scale of automated tools makes any misfires potentially much more damaging.[48]

Given these profound concerns regarding due process and equal protection, it is no surprise that a robust literature, described below, has emerged to suggest paths forward for the regulation of algorithms in government.

Some might assume that *any* system where humans make the ultimate decision would resolve some of the due process issues noted above. But it turns out that the interface between humans and models is a critical ingredient in making human review meaningful. For example, it might seem like a good idea for humans to be responsible for monitoring and correcting errors created by algorithms.[49] But a large

[44] *See, e.g.*, Huq, *supra* note 42, at 644, 666 (discussing the opacity of other minds and the relatively worse performance of human decision-makers); Cary Coglianese & Alicia Lai, *Algorithm vs. Algorithm*, 71 DUKE L.J. 1281, 1311–13 (2022) (discussing the welfare gains from automation).

[45] *See, e.g.*, Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instrument*s, 5 BIG DATA 153, 157 (2017); *see also* Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 823–25 (2020). This is a more general problem. *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 720 (2016).

[46] *See* Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel & Aziz Huq, *Algorithmic Decision Making and the Cost of Fairness*, 2017 PROC. INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 797, 797, https://dl.acm.org/doi/pdf/10.1145/3097983.3098095 [https://perma.cc/UH8T-FZWK]; Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1101, 1124 (2019).

[47] Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. 1ST CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY: PROC. MACH. LEARNING RSCH., 2018, at 1, 10, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf [https://perma.cc/5PB3-ZT2X].

[48] Chouldechova, *supra* note 45, at 162.

[49] *See, e.g.*, Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, *Humans in the Loop*, 76 VAND. L. REV. 429, 474–78 (2023) (discussing the various forms of error

literature (to which we return below) has documented a multitude of cognitive biases that might make that role challenging or impossible for humans to complete successfully.[50] One point of concern is automation bias, when humans overly and selectively believe model predictions that match prior beliefs (themselves derived from potentially illicit sources).[51] Still, the dignitary benefits of knowing that someone is watching the machines might be justified, even if the humans produce fewer gains in accuracy.[52]

Another tack has been to propose system-level oversight interventions. A burgeoning new literature, for example, focuses on the idea of pre-implementation or midstream audits to identify unexpected harms and map algorithms' value for policy objectives.[53] Others have focused on new government agencies or oversight organizations within agencies to supervise implementation.[54] Finally, a simple but powerful proposal is to require randomization when agencies implement algorithmic tools to permit easy comparisons between decisions made with and without new processes.[55]

As the foregoing discussion indicates, the bulk of previous writing on algorithms in government is either situated at a very high level (i.e., at the level of constitutional values) or focused on future reforms that, while compelling, would require either statutory changes or watershed doctrinal changes. By contrast, this paper focuses on the extant

correction that might alleviate due process concerns); Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 1964–65 (2021) (arguing for the right to appeal particular decisions made by automatic decision systems).

[50] Ben Z. Green, *The Flaws of Policies Requiring Human Oversight of Government Algorithms*, 45 COMPUT. & SOC'Y REV. 1, 7–8 (2022); Alon-Barkat & Busuioc, *supra* note 43, at 154–56.

[51] Alon-Barkat & Busuioc, *supra* note 43, at 159–61; Megan T. Stevenson & Jennifer L. Doleac, Algorithmic Risk Assessment in the Hands of Humans 20 (Sept. 29, 2022) (unpublished manuscript), https://dx.doi.org/10.2139/ssrn.3489440 [https://perma.cc/4XGJ-UWFD].

[52] *See, e.g.*, Crootof et al., *supra* note 49, at 481, 490.

[53] *See* Gregory Falco et al., *Governing AI Safety Through Independent Audits*, 3 NATURE MACH. INTEL. 566, 569–70 (2021); Ari E. Waldman, *Power, Process & Automated Decision-Making*, 88 FORDHAM L. REV. 613, 616 (2019); *see also* Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg & Daniel E. Ho, *Outsider Oversight: Designing a Third-Party Audit Ecosystem for AI Governance*, 2022 PROC. ACM CONF. ON AI, ETHICS, & SOC'Y 557, 569, https://arxiv.org/pdf/2206.04737.pdf [https://perma.cc/AY4D-2DFE] (addressing the potential for both pre-implementation and online audits across commercial and government uses of AI). Audits are also related to some of the proposals contained in the literature on AI procurement. *See, e.g.*, Ben Dor & Coglianese, *supra* note 41, at 195 (calling for incorporating guarantees of "conform[ance] to specified standards for ethical AI").

[54] *See, e.g.*, Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83, 106 (2017) (proposing one agency to supervise new algorithms generally); *see also* Engstrom & Ho, *supra* note 2, at 847 (proposing internal agency structures to supervise the adoption of algorithms).

[55] Engstrom & Ho, *supra* note 2, at 849–51 (noting "benchmarking [relative to a random hold-out set] enables decision-makers to directly assess the impact of the AI tool in real time" by providing a "comparison group to smoke out inaccuracies and biases").

doctrinal categories found in administrative law and asks how algorithmic systems will fit.

That is not to say that we are the first to observe the relationship between discretion, algorithms, and the doctrinal treatment of rulemaking in the APA. Nearly fifteen years ago, Danielle Citron observed that algorithms tend to "blur the line between adjudication and rulemaking, confounding the procedural protections governing both systems" and focused on the normative questions of whether a given decision ought to be automated or not and what the implications of that choice might be for due process values.[56] We agree with Citron's starting point. But rather than thinking through the normative questions, this paper aims at understanding just where algorithm-based systems might escape the "procedural protections" designed to cover rulemaking.[57]

Closer to this work is David Freeman Engstrom and Daniel E. Ho's 2020 article, *Algorithmic Accountability in the Administrative State*.[58] Like us, Engstrom and Ho focus on the rule/guidance dichotomy (which we discuss below) as the key issue for algorithmic regulation in extant administrative law.[59] They offer three factors to guide courts and agencies in deciding when an algorithmic system is more like a "rule."[60] Two of those factors are clear on their face: the presence of large distributive consequences and the use of the algorithm for enforcement or adjudication.[61] We agree that these factors are normatively relevant and that the enforcement/adjudication distinction also has great doctrinal significance. But these two factors alone cannot decide every case; a large class of algorithms focuses on adjudication and comes with uncertain distributive consequences, at least ex ante.

Their final factor — the presence of a "genuine exercise of human discretion"[62] — is the most important doctrinal question, but what it means is far from obvious. Our goal here is to give more content to this idea, informed by recent caselaw and computer science research, and thus move toward a more complete picture of how algorithmic development is likely to be shaped by existing doctrine. New developments that we will discuss below suggest that there is an analytical framework that aligns administrative law doctrine with best practices in computer science.

[56] Citron, *supra* note 28, at 1278.

[57] *See id.*

[58] Engstrom & Ho, *supra* note 2.

[59] *Id.* at 806, 845.

[60] *Id.* at 845–46.

[61] *Id.* at 846 (arguing that notice and comment is more appropriate when AI adoption involves "considerable distributive consequences" and when adoption is used for adjudication, rather than enforcement).

[62] *Id.* at 845–46 (arguing that "the more humans remain 'in the loop,' the less notice and comment should be triggered.").

III. Rules, Guidance, and Discretion

This Part introduces the doctrinal question at the heart of this Article: whether an algorithmic system is a rule or mere guidance. It argues that the linchpin of that question is the degree to which line adjudicators retain discretion. But discretion is a famously ambiguous term. We therefore offer our own definition that we argue is well suited to the task of identifying rules in the algorithmic state. Because our definition depends on the idea of updating, we also take the opportunity to offer a brief introduction to Bayesian persuasion, which is the field of microeconomics focused on understanding how information causes rational actors to update their beliefs.

A. Administrative Law Standards

The statutory foundation of the distinction between rulemaking and guidance is 5 U.S.C. § 553(b)(A), the provision of the APA that distinguishes so-called "legislative" rules, which require notice and comment, from "interpretative rules, general statements of policy, or rules of agency organization, procedure, or practice," which do not.[63] Significant literature and caselaw have been developed since the passage of the APA to give meaning to these three exceptions to the notice and comment requirement.

In principle, the main difference between guidance and rules is that rules are binding.[64] As Nicholas Parrillo has written, "Guidance is supposed to leave space for the agency's case-by-case discretion," such that agency officials consider evidence and arguments with a fair and open mind even if they contradict an agency's stated position contained in a guidance document.[65] In contrast, an agency adjudicator who encounters a rule must comply with it, full stop.

The dominant approach for determining when an agency document is a rule that is required to proceed by notice and comment is, by far,

[63] 5 U.S.C. § 553(b)(A).

[64] Nicholas R. Parrillo, *Federal Agency Guidance and the Power to Bind: An Empirical Study of Agencies and Industries*, 36 Yale J. on Reg. 165, 165 (2019). Following Ronald Levin, the discussion that follows largely collapses the two categories of guidance (general statements of policy and interpretations of rules or statutes), and treats guidance under the rules that apply to general statements of policy. Levin, *supra* note 12, at 267 ("[W]e should think of the interpretive rule and policy statement components of § 553(b)(A) as comprising, in a significant and not merely nominal sense, a single exemption — the guidance exemption."). To be sure, courts have articulated special tests to identify binding interpretations, but these are famously circular and difficult to parse. *See, e.g.*, Hoctor v. U.S. Dep't of Agric., 82 F.3d 165, 170 (7th Cir. 1996) (An interpretative rule is one "derived from the regulation by a process reasonably described as interpretation"). Accordingly, we set these aside for the present discussion.

[65] Parrillo, *supra* note 64, at 169.

the "practically binding" test.[66] Under this approach, a document can be binding either because (1) agencies themselves expect the document to be binding on adjudicators and say so, or (2) because it appears to have the effect of binding adjudicators in practice.[67] Courts have approached this second task from both the pre-adjudication perspective, drawing on ex ante facts, and from the perspective of a challenge brought at the enforcement stage once a rule has been implemented. To be sure, courts' willingness to entertain ex ante and ex post challenges, and their emphasis on particular factors, varies to some degree across circuits and time. But at a high level, the test is aimed at getting a functional picture of how a document is used within the agency.[68]

Courts looking at a document ex ante examine the text to determine whether it purports to bind. If the document is "couched in mandatory language, or in terms indicating that it will be regularly applied," it is more likely a rule.[69] While this test is straightforward in principle, it sometimes leads to surprising results — as when the D.C. Circuit held that an EPA press release declaring a change in the use of particular data to study pesticide safety was a rule requiring notice and comment because it used mandatory language that rendered the agency pronouncement "binding as a practical matter."[70] Beyond language, courts might also attend to "whether the substantive effect" of the rule "is sufficiently grave," so that policies with more important real-world effects are more likely to be deemed rules.[71]

In contrast, an ex post inquiry draws on information that becomes available once the implementation of a rule has begun. As the D.C. Circuit held in its watershed opinion *Appalachian Power Co. v. EPA*,[72] if an agency "acts as if a document issued at headquarters is controlling

[66] Levin, *supra* note 12, at 297 (reporting that "lower courts have, for better or worse, embraced" this approach). *But see* Cass R. Sunstein, *Practically Binding: General Policy Statements and Notice-and-Comment Rulemaking*, 68 ADMIN. L. REV. 491, 516 (2016) (concluding that the "practically binding test is an unacceptable departure from any plausible reading of the APA."). Sunstein's view appears to represent the minority of scholarly opinion and, in any event, clearly does not represent the views of most lower courts. Levin, *supra* note 12, at 315 ("[N]o commentator other than Sunstein . . . has subscribed to this view.").

[67] *See* Levin*, supra* note 12, at 293–98 (explaining the two-component "binding norm test").

[68] *Compare* Texas v. United States, 809 F.3d 134, 187–88 (5th Cir. 2015), *aff'd*, 579 U.S. 547 (2016) (adopting a thoroughly functionalist approach to distinguishing rulemaking from guidance), *with* Grace v. Barr, 965 F.3d 883, 906 (D.C. Cir. 2020) (suggesting that evidence of practically binding effect must be combined with evidence of binding language found on the face of the document in question). To some degree, *Texas* itself — which employed a functionalist analysis to find the Deferred Action for Parents of Childhood Arrivals ("DAPA") program illegal — might itself be the cause of some skepticism of the "practically binding" test. *See Texas*, 809 F.3d at 187–88.

[69] General Elec. Co. v. EPA, 290 F.3d 377, 383 (D.C. Cir. 2002) (reasoning "the mandatory language of a document alone can be sufficient to render it binding").

[70] *Id.*

[71] Elec. Priv. Info. Ctr. v. Dep't of Homeland Sec., 653 F.3d 1, 5 (D.C. Cir. 2011).

[72] 208 F.3d 1015 (D.C. Cir. 2000).

in the field," if it "leads private parties" or other government entities to believe they must comply, or if it "bases enforcement actions on the policies or interpretations formulated in the document," then the document is more likely to be deemed a rule.[73] Indeed, evidence of practical deference to the policy articulated in a document can make it a rule even when language on the face of the document suggests that it is not binding.[74] In contrast, if the agency adjudicators exhibit practical evidence of regularly departing from the department's policy, it is more likely to be treated as guidance, notwithstanding apparently mandatory language.[75]

It is important to recognize that the distinction between rules and guidance we have been discussing thus far is of relevance primarily, though not exclusively, in the context of adjudication, rather than enforcement. When agencies make decisions about whether to bring enforcement actions, the *Heckler v. Chaney* doctrine stands for the proposition that the exercise of prosecutorial discretion is typically exempt from judicial review absent explicit Congressional instructions to the contrary.[76] As a result of *Heckler*, documents that bind officials in the exercise of their enforcement discretion are virtually never subject to notice and comment requirements.[77] In recent years, the Fifth Circuit has significantly eroded *Heckler*'s carve-out for prosecutorial discretion, beginning with its decision in *Texas v. United States*, which subjected the Obama Administration's deferred action program for parents of childhood arrivals to notice and comment review.[78] Nonetheless, we assume that algorithms pertaining to enforcement will be largely shielded from notice and comment in the following Sections.

B. Defining Discretion

Despite the importance of discretion in distinguishing rules from guidance, the meaning of that term is ambiguous. Legal philosophers have long wrestled with the question of whether discretion should be

[73] *Id.* at 1021.

[74] McLouth Steel Prods. Corp. v. Thomas*,* 838 F.2d 1317, 1321 (D.C. Cir. 1988) ("More critically than EPA's language adopting the model, its later conduct applying it confirms its binding character.").

[75] Sierra Club v. EPA, 873 F.3d 946, 952 (D.C. Cir. 2017) (finding that a policy was not a rule because, among other things, "EPA's vow to remain flexible was not just talk, as shown by its conduct").

[76] Heckler v. Chaney, 470 U.S. 821, 837–39 (1985); *see also* Citizens for Resp. & Ethics v. Fed. Election Comm'n, 923 F.3d 1141, 1144–45 (D.C. Cir. 2018) (Pillard, J., dissenting from denial of rehearing en banc) (discussing the unique judicial power to review Federal Election Commission enforcement decisions).

[77] *See* Engstrom & Ho, *supra* note 2, at 830–31 nn.107–09 (noting that *Hecker* created a "strong presumption against review" for these documents that can be rebutted only under "narrow circumstances").

[78] *See* Texas v. United States*,* 809 F.3d 134, 171–72 (5th Cir. 2015), *aff'd*, 579 U.S. 547 (2016).

understood as the power to impose one's tastes and preferences at will, or whether discretion implies a thicker, more deliberative process with a stronger claim to rationality.[79] Understood as a mere matter of taste, the word discretion could apply equally to criminal sentencing as to one's preferred wallpaper. Alternatively, discretion might require effortful reasoning, as when judges reason about cases.

This fork in the conceptual road leads to different empirical approaches. If discretion amounts to the power to impose idiosyncratic preferences, then the search for discretion in administrative law would approach the study of judicial ideology in political science. Political scientists assume that every judge has a taste for certain bundles of outcomes that can be projected into a low-dimensional ideological space.[80] But as previous empirical research has shown, low-dimensional ideological preferences may be significantly less predictive of adjudicators' behavior in high-throughput administrative agencies.[81] And even if equating discretion with ideology was not empirically tenuous, it would also be in significant tension with the normative premises of discretion in administrative law, which emphasize the impartiality, expertise, and rationality of decision-makers.

By contrast, the legal positivist view of discretion as the requirement that decision-makers engage in thoughtful, rational consideration is a much closer fit with the traditional vision of adjudicators' roles.[82] It also fits with the empirical observation that adjudicators in real-world agencies are often singularly focused on beating crushing caseloads.[83] In this world, the absence of discretion has more to do with line adjudicators lacking the capacity to consider the underlying issues — and less with constraints on the expression of taste.

[79] *See* Nicola Lacey, *The Path Not Taken: H.L.A. Hart's Harvard Essay on Discretion*, 127 HARV. L. REV. 636, 643–44 (2013). In a recently discovered essay, the famed legal positivist H.L.A. Hart argued for the latter view. *See* H.L.A. Hart, *Discretion*, 127 HARV. L. REV. 652, 656–57 (2013). By contrast, philosophers such as Ronald Dworkin are associated with the former view, equating discretion with mere taste. *See* Ronald Dworkin, *Judicial Discretion*, 60 J. PHIL. 624 (1963).

[80] *See, e.g.*, Andrew D. Martin & Kevin M. Quinn, *Assessing Preference Change on the U.S. Supreme Court*, 23 J.L. ECON. & ORG. 365, 370–71 (2007).

[81] *See* David K. Hausman, Daniel E. Ho, Mark S. Krass & Anne McDonough, *Executive Control of Agency Adjudication: Capacity, Selection, and Precedential Rulemaking*, 39 J.L. ECON. & ORG. 682, 683 (discussing the failure of the Trump administration to appoint immigration judges whose preferences for removing immigrants deviated significantly from those of Obama or Clinton appointees). Indeed, the puzzling failure of traditional ideological categories extends to high-throughput settings outside the administrative state as well. *See* Daniel M. Thompson, *How Partisan Is Local Law Enforcement? Evidence From Sheriff Cooperation with Immigration Authorities*, 114 AM. POL. SCI. REV. 222, 230 (2020) (documenting the null relationship between partisan affiliation and cooperation with immigration authorities).

[82] *See, e.g.*, Jerry L. Mashaw, *Small Things Like Reasons Are Put in a Jar: Reason & Legitimacy in the Administrative State*, 70 FORDHAM L. REV. 17, 25–26 (2001).

[83] *See* Ames et al., *supra* note 11, at 4, 17–19 ("[A]s ALJ production increases, the general trend for decisional quality is to go down.").

In accord with this conceptual definition of discretion, we propose a definition of discretion as effortful updating. On this view, every legal decision begins with a set of prior beliefs. While lay parlance often treats "priors" as the product of purely internal reflection, that is often untrue in the context of law. For example, the law often supplies presumptions (e.g., the presumption of innocence in criminal trials) that must be taken as part of a decision-maker's set of prior beliefs.

A related source of information for adjudicators' priors are previous treatments of the same case. In the context of a computer-based model, a model supplying a final recommendation (e.g., to deny or grant bail) forms part of the decision-maker's set of prior beliefs. To be clear, though, model outputs are far from the only example in this category. For instance, when the Board of Veterans' Appeals reviews a decision to deny benefits, its adjudicators will also see the disposition reached by the Veterans Benefits Administration staff who decided a veteran's entitlement in the first instance, and indeed must defer to that initial decision to the extent it favors the veteran.[84]

Against this backdrop, we understand "discretion" to mean adding to this set of prior beliefs by expending effort to perceive, gather, or interpret information about a given case. For example, an immigration judge exercises discretion by comparing an algorithm's recommendation about whether to grant bail against her own judgment about the immigrant's likely recidivism, given her interaction with the immigrant and a holistic examination of the circumstances of the case. In contrast, an immigration judge would fail to exercise discretion if she applied a uniform decision rule, such as always deferring to the algorithm's initial bail recommendation, with no regard whatsoever to the factual circumstances of a particular case. In short, discretion means updating one's prior beliefs by integrating them with information one perceives.

C. Bayesian Persuasion, Information Design, and Human-Computer Interaction

In defining the exercise of discretion as the practice of updating prior beliefs, we are deliberately alluding to the idea of updating that is common in the study of persuasion. In Part IV, below, we draw from a growing economics literature that studies "Bayesian persuasion," that is, the way information causes rational listeners to update their prior beliefs.[85]

[84] *See* 38 C.F.R. § 20.801 (2019) ("Any findings favorable to the claimant as identified by the agency of original jurisdiction in notification of a decision . . . are binding on all agency of original jurisdiction and Board of Veterans' Appeals adjudicators, unless rebutted by evidence that identifies a clear and unmistakable error in the favorable finding.").

[85] For a review of this literature, see, for example, Emir Kamenica, *Bayesian Persuasion and Information Design*, 11 ANN. REV. ECON. 249 (2019).

To help readers unfamiliar with that literature, this Section offers a quick primer on the basic setup.

In their seminal paper on Bayesian persuasion, Emir Kamenica and Matthew Gentzkow begin with two stylized characters: a prosecutor and a judge, the latter of whom is imagined as the sole decision-maker about an accused person's criminal guilt.[86] The prosecutor has several options for how to present her evidence to the judge. For example, she controls what evidence she puts on, the order in which she does so, the strength of her language, and so forth. Kamenica and Gentzkow focus on understanding how the prosecutor should weigh those options in order to maximize her chances of convincing the judge of the defendant's guilt. Here, we will swap out the prosecutor/judge characters for an algorithm, which "decides" how to present information and recommendations to a line officer, who is presumed to act rationally. (You can also think of the designer behind the algorithm, or the policymaker that controls the algorithm, as the prosecutor, if you like.) While our setting is different than Kamenica and Gentzkow's, the focus is the same: We are interested in understanding how the algorithm's strategy for presenting data might make the human more likely to conform to the algorithm's recommendations.[87]

One more piece of terminology: If the algorithm recommends an action that it would like the line officer to take, this is known as a "direct" persuasive scheme. This terminology, articulated by Professor Haifeng Xu, outlines a persuasion strategy in the sense that the algorithm is rewarded if it causes the line officer to update her beliefs such that her preferred action is the one recommended by the algorithm.[88]

We will also discuss a number of studies that reveal potentially irrational human behaviors resulting from confrontations with algorithmic recommendations. These studies are centered in the field of human-computer interaction ("HCI"). This field of research emerged in the 1980s "as a specialty area in computer science embracing cognitive science and human factors engineering" and has since expanded into a diverse, interdisciplinary field.[89] In relation to algorithms, HCI has examined how humans interact with the algorithms and, in particular,

[86] Emir Kamenica & Matthew Gentzkow, *Bayesian Persuasion*, 101 AM. ECON. REV. 2590, 2590 (2011).

[87] Of course, in reality, humans are not rational, which is why we supplement the Bayesian persuasion literature with an examination of Human-Computer Interaction literatures.

[88] *See* Haifeng Xu, *Algorithmic Persuasion*, 2018 ACM CONF. ON ECON. & COMPUTATION (June 18, 2018), https://www.haifeng-xu.com/information-ec18/part3-persuasion.pdf [https://perma.cc/B9XJ-LLMU]; *see also* Kamenica & Gentzkow, *supra* note 86, at 2590 (illustrating such a strategy).

[89] INTERACTION DESIGN FOUND., *THE ENCYCLOPEDIA OF HUMAN-COMPUTER INTERACTION* (2d ed. 2014), [https://perma.cc/R8YG-UFWT].

what factors lead to automation bias.[90] Automation bias occurs when humans blindly accept algorithmic recommendations, even when the algorithm is error-prone.[91] Others have noted that humans selectively accept algorithmic recommendations to reaffirm their biases.[92] HCI studies then follow up to break down what aspects of algorithm or interface design lead to such behaviors. We offer several examples below.

IV. A Functionalist Framework for Distinguishing Algorithmic Rulemaking from Algorithmic Guidance

In deciding whether a document is a rule that practically binds an agency, or guidance that does not, courts look to both ex ante factors, like the presence of mandatory language, and ex post factors, like the on-the-ground treatment of a document as binding.[93] This Part presents the heart of our analysis, mapping how each of those lines of analysis might play out when the procedure in question involves an algorithm. It bears emphasizing that this Part assumes that courts remain focused on how a guidance document will work in practice — in short, that the "practically binding" test we discuss in Section III.A remains good law.

None of the factors we discuss is independently necessary for finding an algorithm to be a rule. Rather, an accumulation of factors increases the probability that an algorithm is "practically binding" and thus subject to the notice and comment process. We start by examining the ex ante factors that we think are most likely to be dispositive. These rely on structural arguments informed by Bayesian persuasion and human-computer interaction studies. We then consider ex post factors which tend to require use of data and causal evidence. Because of the significant challenges associated with retrospective data analysis in legal settings, we are more pessimistic about the potential for courts to identify rules by relying on ex post empirical data, at least without more systematic data collection on the part of agencies. In Part V, we consider the potentially salutary effect of our framework on agency incentives.

[90] *See, e.g.*, Maria De-Arteaga, Riccardo Fogliato & Alexandra Chouldechova, *A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores*, 2020 Proc. CHI Conf. on Hum. Factors in Computing Sys., Apr. 2020, at 1, 2, https://arxiv.org/pdf/2002.08035.pdf [https://perma.cc/HH55-7W5S%5D].

[91] *Id.*

[92] *See, e.g.*, Alon-Barkat & Busuioc, *supra* note 43, at 154 (noting "decision-makers' *selective adherence* to algorithmic advice"); Megan T. Stevenson & Jennifer L. Doleac, Algorithmic Risk Assessment in the Hands of Humans 25–26 (Apr. 22, 2021) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440 [https://perma.cc/Y568-HWT5].

[93] *See* Levin, *supra* note 12, at 290–300 (describing courts' examination of mandatory language and binding effects in practice).

A. Ex Ante Factors

Consider a line adjudicator at an agency faced with some decision who receives assistance — a "signal" — from a computer-based algorithm deployed by the agency. That signal might be either a recommendation as to the final disposition of the decision (e.g., the algorithm outputs a suggested bail amount) or a recommendation as to some subcomponent of the decision (e.g., how to find the right cases and evidence). This Section considers what facts about the way the signal is structured, or about the context in which the signal is received, could help predict whether it will be "practically binding," that is, so persuasive as to induce compliance.

We approach that task using the framework of the Bayesian persuasion games introduced in Section III.C. As we note there, Bayesian persuasion analyses include two roles: a "sender" who aims to persuade, and a "receiver" who perceives the sender's signals and wishes to make an optimal decision. We treat algorithms as senders: They are the source of a signal that can persuade. Line adjudicators are the receivers: They receive signals and try to integrate them with everything else they know, with the overall objective of trying to reach an accurate decision. To make the crosswalk between our discussion and the literature more direct, we simply refer to "senders" and "receivers." This framework, although somewhat stylized, allows us to draw on the Bayesian persuasion literature to generate theoretical predictions for when models should be most persuasive. In the final part of this Section, we also consider how contextual factors outside the model itself could shift users' tendency to defer to the algorithm's outputs. Note that throughout we will also reference the human-computer interaction literature, which empirically provides clues about what design features would make the algorithm more or less binding — particularly when humans behave irrationally.

B. Incentives and Ease of Independent Evidentiary Searches

One critical variable is the degree to which an adjudicator can easily access auxiliary evidence against which the algorithm's prediction may be evaluated. In the context of administrative adjudications, "auxiliary evidence" might mean parts of the factual record; in an immigration appeal, for example, this might mean seamless access to data on the claimant's country of origin or a freeform description of the facts of the case. The easier auxiliary evidence is to access, the more informative it is; and the clearer its structure, the less likely an adjudicator is to defer to it. We first provide intuition for this claim and then present the empirical and theoretical results substantiating it.

Consider the extreme case in which an adjudicator simply receives a bare risk score — that a criminal defendant will recidivate, say — with no access to any other information save that bare prediction. If the risk prediction is the only relevant information available to the decision-maker, then they must rely completely on the prediction. This hypothetical illustrates the total absence of discretion: a rational decision-maker would have no room for deliberation, and the only possible variation would come from the imposition of idiosyncratic tastes.[94]

This extreme example highlights the fact that exercising "discretion" in the sense of engaging in some deliberation requires that the decision-maker have access to signals beyond the algorithm itself. This observation highlights the importance of the often-antiquated systems — sometimes even including voluminous paper records — through which agency adjudicators frequently access record materials.[95] When access to the record is poor, reliance on algorithmic recommendations may increase.

This intuitive claim is supported by empirical evidence. One study considers an algorithm used to assist with child maltreatment hotline screening decisions, where the algorithm produced a risk score (akin to recidivism predictions in the criminal or immigration context).[96] As one would expect, the algorithm's recommendations were generally persuasive and changed the outcomes of cases.[97]

But the user interface, which made it easy to access underlying information about each case, prevented that reliance from being blind.[98] For instance, in the middle of deployment, the risk assessment algorithm was erroneously updated such that users inadvertently received incorrect risk scores.[99] Surprisingly, the human end users spotted and avoided these errors. This is all the more striking because users faced a mild cost for deviating from the system's recommendations in that they had to seek managerial approval to override the algorithm.[100] In other words, the authors speculated that ease of access to auxiliary signals helped overcome the cost of managerial approval to correct the

[94] *See supra* Section III.B (discussing the meaning of discretion).

[95] For a discussion of the state of electronic case management systems in administrative agencies, see FELIX F. BAJANDAS & GERALD K. RAY, ADMIN. CONF. OF THE U.S., IMPLEMENTATION AND USE OF ELECTRONIC CASE MANAGEMENT SYSTEMS IN FEDERAL AGENCY ADJUDICATION 31–45.

[96] De-Arteaga et al., *supra* note 90, at 3.

[97] *Id.* at 5–6.

[98] *Id.* at 9 ("A key contributing factor is that throughout the process call workers continued to have access to not only the referral calls but also the administrative data system. This provided a different view of the case than what was being pulled into the risk score calculation. In particular, even when inputs related to past child welfare history were being miscalculated in real time, workers would still have access to the correct information in the data system.").

[99] *Id.* at 4.

[100] *Id.* at 7.

algorithmic error and maximize the chance of being correct. In short, auxiliary information is key to reducing overdependence on algorithmic outputs.

This empirical evidence also matches more formal treatments of the same problem. In a paper by Matysková and Montes, a receiver (who occupies the role of adjudicator in our setting) can pay some cost to get additional information after observing the sender's (algorithm's) signal, but before taking an action.[101] In such a scenario, "[w]hen the receiver faces a lower cost of information, her 'threat' of gathering independent information increases, thus decreasing the sender's power to persuade."[102] Just as was true in the real-world example of the child maltreatment hotline, this theoretical result suggests that access to additional information increases the likelihood that an officer will exercise their discretion, even when doing so comes with a cost like seeking supervisor permission to override the algorithm.

Other theoretical work has emphasized the potential value of balancing algorithmic recommendations with alternative views. One suggestion is that humans are likely to make better overall decisions when they draw on information to which the algorithm does not have access and combine it with the algorithm's recommendation.[103] For instance, one might imagine that immigration officers who combine an algorithmic review of the record with in-person judgments of credibility would perform better than those who merely review the same set of documents available to the algorithm.

A related approach is to accompany predictions with auxiliary evidence likely to dissent from the algorithm's conclusion.[104] For example, a system could identify experts likely to disagree with a particular recommendation and then recommend that users seek input from those experts in addition to the algorithm's recommendation.[105] Although efforts to implement similar approaches have proven challenging,[106] the

[101] *See* Ludmila Matysková & Alfonso Montes, *Bayesian Persuasion with Costly Information Acquisition*, 211 J. ECON. THEORY 1, 3–4 (2023). However, the lower cost of information may hurt the receiver as the persuasive sender might change their signal to reduce the amount of information provided in the equilibrium path.

[102] *Id.* at 1.

[103] *See* Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing & Gerhard Satzger, *On the Effect of Information Asymmetry in Human-AI Teams*, 2022 CHI CONF. ON HUM. FACTORS IN COMPUTING SYS. 1–2, https://arxiv.org/pdf/2205.01467.pdf [https://perma.cc/3DN8-G4UA]. Although Hemmer et al. articulate this theoretical explanation, note that their empirical results do not positively support such an effect.

[104] Matysková & Montes, *supra* note 101, at 3–4.

[105] Maria De-Arteaga, Alexandra Chouldechova & Artur Dubrawski, *Doubting AI Predictions: Influence-Driven Second Opinion Recommendation*, ACM CHI 2022 WORKSHOP ON TRUST & RELIANCE IN AI-HUM. TEAMS 1–2, https://arxiv.org/pdf/2205.00072.pdf [https://perma.cc/BA2R-QRPD].

[106] *See generally* Daniel E. Ho & Lisa Larrimore Ouellette, *Improving Scientific Judgments in Law and Government: A Field Experiment of Patent Peer Review*, 17 J. EMPIRICAL

more general idea of actively serving evidence that contradicts an algorithm's recommendation is a compelling extension of the auxiliary evidence idea.

In all, rational adjudicators whose goal is to find the correct answer are less likely to defer to algorithms and are more likely to exercise discretion when their access to alternative sources of information is seamless. To be sure, getting the right answer is not always adjudicators' primary objective. Agency officials may simply seek to minimize effort, which would lead to an optimum of rubber-stamping the algorithm's decision since this is the lowest effort action.[107] As such, the officer's utility must include some incentive to pursue additional information. In the case of the De-Arteaga, Fogliato, and Chouldechova study, there was likely an implicit incentive to ensure the accuracy of the system: the humans in the loop likely genuinely cared about the safety of children and wanted to ensure that calls were properly assessed. A judge evaluating such a system may think about whether there is such an implicit incentive — and if not, whether there was an artificial incentive like requiring that a short writeup be provided on the evidence that led to a decision. We return to the possibility of alternative objective functions in Section IV.D, below.

C. The Cost of Deviation

Perhaps a more obvious ex ante factor that would indicate a more rule-like algorithm is whether the agency places a cost on disagreement with the algorithm's recommendation. The cost might be something as simple as requiring manager approval for deviation.[108] Or it could require writing an additional report for why there was a deviation and even jeopardize employment for too many deviations. The stronger the cost, the less likely a line officer is to deviate from the algorithm's recommendation, and the more likely the officer is to abandon their own discretionary preferences.

Imposing costs on officials who disagree with an algorithm is not a theoretical scenario. For example, one element of the recidivism

LEGAL STUD. 190 (2020) (documenting an experiment in which experts were solicited to provide specific feedback on patent applications in hopes of improving adjudicators' judgments about the originality of inventions). The intervention "increased examiner search efforts and citations to non-patent literature and reduced the propensity to initially grant the application," but was extremely expensive to implement. *Id.* at 191.

[107] *See, e.g.*, Ames et al., *supra* note 11, at 59–66 (documenting the human factors including the agency's culture, personnel, and political environment, that led the Board of Veterans' Appeals to consistently misstate error rates over many years).

[108] For example, in a national security context, when deciding on whether to execute a drone strike, military personnel were required to seek White House approval if the Collateral Damage Estimation algorithm suggested their strike would yield over thirty civilian casualties. *See* John R. Emery, *Probabilities Towards Death: Bugsplat, Algorithmic Assassinations, and Ethical Due Care*, 8 CRITICAL MIL. STUD. 179, 188–89 (2022).

prediction algorithm used by ICE was that line officers were required to "*provide reasons for any disagreement*" with the algorithm's prediction.[109] This kind of system design, on the margin, will likely push decision-makers towards greater compliance and less discretion.

Of course, imposing costs for disagreement with the algorithm does not always induce blind compliance. For example, in the child maltreatment example we reference above, imposing a modest cost on disagreeing with the algorithm did not stop users from doing so when the algorithm began producing irrational results.[110] The authors of that study suspected that the easy availability of auxiliary information outweighed the cost of deviation, which demonstrates the potential interactions between factors.[111] Similarly, theoretical work on Bayesian persuasion has taken for granted that receivers (i.e., adjudicators) may be willing to pay some cost to acquire reliable information.[112] In sum, the presence of other factors pushing decision-makers towards discretion may outweigh the deterrent effects of deviation costs.

The Bayesian persuasion literature also gives us the converse lesson: Some types of resistance strategies can encourage deviation.[113] That is, under certain conditions, an agent may be more likely to question the signal they receive if they are required to pay some cost for agreeing with the signal, where that cost is determined randomly and before decision time.[114] To make that idea concrete, consider the Department of Homeland Security's RCA algorithm, which has been used since 2012 "to recommend whether to detain or release migrants pending resolution of removal charges."[115] Imagine if ICE agents were required to write reports in a random selection of fifty percent of the cases where they chose to comply with the RCA's score, but never had to do so when disagreeing with the RCA's recommendation. This is, of course, the converse of the truth: agents were required to "provide reasons for any disagreement" with the algorithm.[116] In simplified models, the persuasion literature suggests that resistance strategies motivate the line officer to comply only if the signal is truly informative and improves their decision. In short, imposing costs on accepting algorithmic recommendations can prevent users from surrendering their discretion to irrational persuasion.

[109] Koulish & Evans, *supra* note 32, at 14 (emphasis added).

[110] *See* De-Arteaga et al., *supra* note 90, at 6–8.

[111] *Id.*

[112] *See* Matysková & Montes, *supra* note 101, at 4 (discussing the assumed cost function facing the receiver when they choose to acquire information).

[113] *See* Elias Tsakas, Nikolas Tsakas & Dimitrios Xefteris, *Resisting Persuasion*, 72 ECON. THEORY 723, 732 (2021).

[114] *Id.* at 723 (arguing that "*stochastic resistance strategies* can increase both the informativeness of the signal and the Receiver's payoffs.").

[115] Koulish & Evans, *supra* note 32, at 3–4.

[116] *Id.* at 14.

Finally, making noncompliance costly may also have perverse consequences beyond reliance on the algorithm. One possibility is that adjudicators may try to sidestep prescribed procedures entirely. For example, under the Bush Administration, when deciding on whether to execute a drone strike, military personnel were required to seek White House approval if a Collateral Damage Estimation ("CDE") algorithm suggested their strike would yield over thirty civilian casualties.[117] But when personnel did not wish to incur this cost, they would purposefully tweak the payload used to ensure that it did not meet this casualty threshold.[118] Effectively, the bounds of their discretion were defined by the CDE threshold.

D. Explainability and Informativeness

Consider two algorithms. The first uses a direct persuasive scheme: It recommends an action that it wants the adjudicator to take. There is little cognitive effort required for the adjudicator to simply accept the recommended action; it merely requires the click of a button. By contrast, consider a model that identifies factors relevant to a particular recommendation — or even one that offers reasons without drawing a bottom-line conclusion. The Bayesian persuasion literature refers to this difference as a difference in the structure and value of the signal that the model sends to the receiver.

In this Section, we describe how the structure of signals, in particular the presence of an explanation, might affect reliance on an algorithm's outputs — especially when combined with other factors leading to greater dependence. Two different literatures have addressed how the structure of a signal affects persuasion. In the Bayesian persuasion approach, officers behave rationally based on the information they receive.[119] In the empirical literature on human-computer interaction, the key question is how design decisions affect human decisions.[120] Both provide us with insights into how the presence or absence of explanation might make an algorithmic system more closely resemble a rule.[121]

[117] *See* Emery, *supra* note 108, at 188–89.

[118] *Id.* at 189.

[119] Kamenica & Gentzkow, *supra* note 86, at 2592.

[120] *See, e.g.*, Stephan Diederich, Alfred Benedikt Brendel, Stefan Morena & Lutz Kolbe, *On the Design of and Interaction with Conversational Agents: An Organizing and Assessing Review of Human-Computer Interaction Research*, 23 J. ASS'N FOR INFO. SYS. 96, 97 (2022).

[121] In this sense, the discussion that follows is directly related to the pervasive concern with opacity in algorithmic governance documented above in Section II.B. But even that discussion leaves out many other contributions on this subject. *See, e.g.*, Hannah Bloch-Wehba, *Access to Algorithms*, 88 FORDHAM L. REV. 1265, 1269 (2020) ("The greater the decisional power of the technology, the higher the risk that arbitrary or opaque decisions might evade explanation." In turn, this arbitrariness raises potential "credibility, fairness, and due process implications") (citations omitted); Sylvia Lu, Note, *Data Privacy, Human Rights, and*

The HCI literature presents a mixed picture of how explanation affects an algorithm's power to persuade users. At a high level, explanation often improves performance, and as we have already alluded to, it may have important normative benefits as well.[122]

While the empirical literature is growing, existing evidence in the HCI literature suggests that explainability has a more ambiguous role in the guidance-rule distinction.[123] Users might be *more* persuaded to accept an algorithmic action when presented with an explanation.[124] What's worse, officers might anchor to the explanation, framing their own analysis in the context of this explanation, potentially abandoning their own preferences in favor of those proposed by the algorithm.[125]

But specifics matter: Careful calibration of the way information is presented can have dramatic effects on how users respond. One study focused on the complexity of explanations. It found significant decreases in deference and automation bias, so long as users were presented with simplified explanations of the agent's reasons for a recommended action.[126] But if the amount of information and

Algorithmic Opacity, 110 CAL. L. REV. 2087, 2101–07 (2022) (documenting the data privacy and democratic values compromised by algorithmic opacity); Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 4 (2019) (critiquing the "black box" character of learning algorithms); Ashley Deeks, Essay, *The Judicial Demand for Explainable Artificial Intelligence*, 119 COLUM. L. REV. 1829, 1841–42 (2019) (noting that explainable AI may help assuage concerns about "granting opaque algorithmic decisionmaking a 'presumption of regularity.'") (citations omitted).

[122] *See* Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl & Michael Vössing, *A Meta-Analysis on the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making*, 2022 PROC. AAAI/ACM CONF. ON AI, ETHICS, & SOC'Y 617, 622 figs.2 & 3, https://dl.acm.org/doi/10.1145/3514094.3534128 [https://perma.cc/SD2N-AL3M].

[123] *See, e.g.*, Johannes Jakubik, Jakob Schöffer, Vincent Hoge, Michael Vössing & Niklas Kühl, *An Empirical Evaluation of Estimated Outcomes as Explanations in Human-AI Decision-Making*, 2022 JOINT EUR. CONF. ON MACH. LEARNING & KNOWLEDGE DISCOVERY IN DATABASES 353, 363, https://arxiv.org/pdf/2208.04181.pdf [https://perma.cc/4P99-3RQC]; *cf.* Amit Sharma & Dan Cosley, *Do Social Explanations Work? Studying and Modeling the Effects of Social Explanations in Recommender Systems*, 4 PROC. TWENTY-SECOND INT'L CONF. ON WORLD WIDE WEB 1133, 1137–38 (2013), https://arxiv.org/pdf/1304.3405.pdf [https://perma.cc/7XLW-43YM] (finding that, for social explanations, the form of explanation matters and different modalities persuade users to different extents).

[124] *Id.*

[125] *See, e.g.*, Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy & Klaus Mueller, *Explainable Active Learning (XAL) Toward AI Explanations as Interfaces for Machine Teachers*, PROC. ACM ON HUM.-COMPUT. INTERACTION, 2020, at 1, 14, https://dl.acm.org/doi/10.1145/3432934 [https://perma.cc/TV2B-URP8].

[126] Julia L. Wright, Jessie Y.C. Chen, Michael J. Barnes & Peter A. Hancock, *The Effect of Agent Reasoning Transparency on Automation Bias: An Analysis of Response Performance*, 2016 8TH INT'L CONF. ON VIRTUAL, AUGMENTED & MIXED REALITY 465, 475, https://link.springer.com/chapter/10.1007/978-3-319-39907-2_45 [https://perma.cc/RC68-9E5G] ("When the transparency of agent reasoning was increased to its highest level, complacent behavior increased to nearly the same level as in the no-reasoning condition. This pattern of results indicated that while access to agent reasoning in a decision-supporting agent can counter automation bias, too much information results in an out-of-the-loop (OOTL) situation and increased complacent behavior.").

transparency increased, users were overwhelmed and more likely to blindly defer to the algorithm.[127] Others have pointed to the quality of the explanations. When presented with inaccurate ("low-veracity") explanations for correct recommendations, users were more inclined to disagree with the algorithm, although the algorithm's bottom-line result was right.[128] High-veracity explanations pushed users towards correct answers.[129]

Like at least some of the empirical HCI literature, the Bayesian persuasion literature suggests that, in theory, the presence of explanations should make decision-makers more likely to defer to an algorithm's recommendations.[130] For example, the presence of an explanation might allow a human to spot an incorrect inference drawn from a particular piece of evidence. In many ways, supplementing a recommendation with evidence is analogous to the strategy of a prosecutor in the classic Bayesian persuasion game: Both the prosecutor (sender) and judge (receiver) know in advance that the prosecutor's objective is to persuade the judge to convict, but the prosecutor sequentially provides evidence that biases the rational judge's information environment in favor of that outcome.[131] Here, the end user of an algorithm knows what the algorithm "aims" to persuade them of; that is simply the algorithm's bottom-line recommendation. But the algorithm can offer arguments in favor of that position to bring the user around to that outcome. As in the classic persuasion setting, if the judge does not have her own preference and she is fully informed by the algorithm and its explanation, then the outcome would be in line with the experimental evidence. In short, explanations may be more likely to convince a rational judge — or a rational user — of an algorithm.

But again, this assumes that judges do not have a preference, and that they do not have access to external information (or an incentive to pursue that information to identify errors). Of course, such a simplistic regime would tend to be more rule-like when coupled with persuasive explanations. Crucially, the interaction of the explanation with other

[127] *Id.*

[128] Mahsan Nourani, Chiradeep Roy, Tahrima Rahman, Eric D. Ragan, Nicholas Ruozzi & Vibhav Gogate, *Don't Explain Without Verifying Veracity: An Evaluation of Explainable AI with Video Activity Recognition*, 1 ACM TRANSACTIONS ON COMPUT.-HUM. INTERACTION 1, 19 (2020).

[129] *Id.* at 18.

[130] *See, e.g.*, Finale Doshi-Velez & Been Kim, Towards a Rigorous Science of Interpretable Machine Learning 1, 3 (Mar. 2, 2017) (unpublished manuscript), https://arxiv.org/pdf/1702.08608.pdf [https://perma.cc/V7JC-PAH8] ("[I]f the system can explain its reasoning, we then can verify whether that reasoning is sound with respect to . . . other desiderata — such as fairness, privacy, reliability, robustness, causality, usability and trust . . . ."); Deeks, *supra* note 121, at 1833 ("[S]hedding light on how an algorithm produces its recommendations can help address the other two critiques, by allowing observers to identify biases and errors in the algorithm.").

[131] Recall that the prosecutor sends a signal and the judge receives it. *See* Kamenica & Gentzkow, *supra* note 86, at 2593–94.

factors is important in determining how binding an algorithm's recommendations would be in practice.

One scenario that we have not yet addressed arises when the algorithm is wrong and the human user needs to detect the algorithm's mistake. In some sense, the whole point of ensuring humans remain in the driver's seat is to catch these kinds of mistakes.[132] Will they? We can update the Bayesian persuasion game slightly to model that scenario. In this new setup, the human is tasked with finding mistakes in the algorithm's predictions based on the evidence shown by the algorithm. A 2022 paper by Ederer and Min addresses a similar setting and asks whether lie detection capability on the part of the receiver would change the human's likelihood of accepting the algorithm's recommendation.[133] They find that the receiver's overall performance (framed as their payoff) increases if the receiver's cost of detecting mistakes is sufficiently low.[134] That is, if the receiver has to invest a great deal into detecting model mistakes, then errors are likely to degrade system performance.

Think back to our explanation game. The algorithm makes mistakes or lies at some rate, providing explanations or false recommendations. If the officer is experienced enough, their rate of detection may be sufficiently high such that the explanations are useful for catching algorithmic errors. But if the officer is inexperienced, their rate of detection might be low and they will end up over-relying on the algorithm.

Another variant involves changing the receiver's access to a final model recommendation. After all, key to any persuasive effects found in the studies above is that the user sees the model's bottom-line take on whatever task they are engaged in. That might create an anchoring effect around the model's judgment — causing explanations to persuade the user to accept the final model's prediction regardless of the correctness of that prediction.[135] What happens when the algorithm omits its final recommendation?

[132] *See, e.g.*, Sofia Ranchordas, *Empathy in the Digital Administrative State*, 71 DUKE L.J. 1341, 1349–79 (2022).

[133] Florian Ederer & Weicheng Min, *Bayesian Persuasion with Lie Detection* 3 (Nat'l Bureau of Econ. Rsch., Working Paper No. 30065, 2022).

[134] *Id.* at 3, 18 (assuming that the lie detection technology is "sufficiently reliable," any further increase in the lie detection probability causes the receiver's equilibrium payoff to increase with this probability).

[135] *See* Ghai et al., *supra* note 125, at 14 (finding that explanations made users more likely to accept a model's final prediction even when the final prediction was incorrect); Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar et al., *Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance*, PROC. CHI CONF. ON HUM. FACTORS COMPUTING SYS., May 2021, at 1, 1, https://dl.acm.org/doi/10.1145/3411764.3445717 [https://perma.cc/32TK-DDH6] (similarly finding that explanations increased the chance the users would accept the model recommendation, even when the recommendation was wrong).

Without a recommendation to lean on, officers have to examine evidence that might be important in making a determination but are not provided with an explanation of how to piece those features together. AI systems that suggest relevant citations might fall into this category: they do not ultimately suggest an outcome, but rather point toward relevant inputs to that decision.[136]

In this setting, we might look to the studies of Bayesian persuasion in which the sender can only send limited information or cannot fully describe a recommended action.[137] A paper by Aybas and Turkel, for example, addresses a theoretical context in which an advertiser is prevented from providing full information about their product to consumers by a regulator, such as when this hypothetical regulator seeks to limit the targeting capability of advertisers to improve consumers' welfare.[138] Aybas's and Turkel's work, when translated to our setting, suggests that the more pieces of evidence an algorithm can provide to the officer, the more chances there are to persuade.[139] And the more uncertain the officer is (e.g., the officer does not have other information to look to or is not well-trained to conduct an independent investigation), the more persuasive those pieces of evidence will be.

We might extrapolate from Aybas's and Turkel's research that preventing the algorithm from showing a final recommendation would reduce human dependence on the algorithm's recommendation. Under our framework, that would make the algorithm less rule-like. The APA, and potentially other legislation, impose one constraint on that principle by regulating the kinds of information adjudicators must consider before making a decision.[140]

To see that principle in action, consider two recent cases addressing the Department of Homeland Security's Risk Classification Assessment algorithm.[141] In *Fraihat v. ICE*,[142] a federal district court found

[136] For an example of such a system, see generally Huang et al., *supra* note 23.

[137] *See* Yunus C. Aybas & Eray Turkel, Persuasion with Coarse Communication (Oct. 29, 2019) (unpublished manuscript), https://arxiv.org/abs/1910.13547 [https://perma.cc/DG3M-6RZB]; *see also* Elias Tsakas & Nikolas Tsakas, *Noisy Persuasion*, 130 GAMES & ECON. BEHAV. 44 (2021); Shaddin Dughmi, David Kempe & Ruixin Qiang, *Persuasion with Limited Communication*, 2016 PROC. ACM CONF. ON ECON. & COMPUTATION 663, https://dl.acm.org/doi/10.1145/2940716.2940781 [https://perma.cc/64NW-X25N]; Shota Ichihashi, *Limiting Sender's Information in Bayesian Persuasion*, 117 GAMES & ECON. BEHAV. 276 (2019).

[138] Aybas & Turkel, *supra* note 137, at 2.

[139] *See id.* at 17 (finding that additional signal is always more valuable for the Sender — in our case signals can be pieces of evidence shown to an officer the Sender is the algorithm).

[140] Motor Vehicles Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co.*,* 463 U.S. 29, 43 (1983) (holding that an agency must "examine the relevant data and articulate a satisfactory explanation for its action including a rational connection between the facts found and the choice made") (internal quotation marks omitted).

[141] The RCA algorithm is a point-based algorithm, taking into account a range of factors about a particular case drawn from ICE data files to make a bail determination recommendation. *See* Koulish & Evans, *supra* note 32.

[142] 445 F. Supp. 3d 709 (C.D. Cal. 2020).

that the medical questionnaire used as input to the RCA did not sufficiently account for the vulnerabilities of detainees to COVID-19 in making its release recommendations.[143] Because the RCA had failed to consider relevant information, the plaintiffs were found to have stated a viable claim under Section 504 of the Rehabilitation Act of 1973 to warrant issuance of a preliminary injunction.[144] And in *Ramirez v. ICE*,[145] another federal district court found that officers' failure to consider detainees' age in making release determinations, due to their overreliance on the RCA, violated the principle that minors must be detained in the "least restrictive setting available after taking into account [their] danger to self, danger to the community, and risk of flight."[146] Because the algorithm was incapable of incorporating the evidence that was legally required to be factored into a final decision (i.e., the status of detainees as minors), decisions based on the algorithm were necessarily arbitrary. Both of these cases illustrate that algorithms can enable heavy-handed regulatory regimes by excluding legally relevant information.

While both *Fraihat* and *Ramirez* speak to the importance of making decisions on the basis of all legally required information, they are distinguishable from the kinds of discretion-preserving algorithms we mention above. Both cases involved officer reliance on a bottom-line recommendation that relied on a deficient set of information. They did not address a world in which the entire purpose of the algorithm was to surface the most informative pieces of evidence or the most important legal sources for the adjudicator to then incorporate into a considered decision. For the reasons we describe above, an algorithm focused on that kind of research-assistant role would be far less likely to impinge on an adjudicator's discretion in a way that would invoke the APA.

Thus far, we have focused on the structure of the signal that the model sends to an adjudicator. Needless to say, the other factors we discuss in Part IV are likely to interact with the signal to shape discretion. The personal characteristics of adjudicators matter too. The HCI literature, for instance, suggests that users' self-confidence and degree of experience might influence deference to the algorithm's recommendations.[147] Though the legal dimension of an algorithmic system usually cannot be conditioned on the identities of the staff who use it, it is worth bearing in mind that the structural explanation has additional considerations.

[143] *See id.* at 728, 748.

[144] *Id.* at 747–50.

[145] 471 F. Supp. 3d 88 (D.D.C. 2020).

[146] *Id.* at 92 (quoting 8 U.S.C. § 1232(c)(2)(B)).

[147] *See, e.g.*, Jennifer-Marie Logg, *Theory of Machine: When Do People Rely on Algorithms?* 29, 50 (Harv. Bus. Sch., Working Paper No. 17-086, 2017), https://dash.harvard.edu/handle/1/31677474 [https://perma.cc/QVY9-H6GL].

To sum up, recommending evidence, citations, or other inputs to a final decision rather than a bottom-line decision is more likely to preserve an officer's discretion, and is thus less likely to be an APA rule. One intuitive way to understand this argument is that such an algorithm would leave the adjudicator with several more reasoning steps between its output and a final decision.

To put things more starkly, guidance-like explanations truly aid the officer in making a decision, while rule-like explanations persuade the officer to trust the algorithm. Distinguishing the two can be difficult in some cases, but is nonetheless possible. For example, we might consider the Social Security Administration's Insight system, which "enables adjudicators to check draft decisions for roughly 30 quality issues."[148] This system flags errors that lead to a successful appeal — such as leaving a claim unaddressed — as inputs to a decision. It does not persuade the adjudicator on how to evaluate the bottom-line claim.

E. The Optimization Objective

Courts should also pay close attention to the outcome that a model was trained or purchased to optimize for — that is, the optimization objective. A particularly important distinction is between models that aim primarily to generate maximally *persuasive* suggestions and those aimed at producing maximally *accurate* suggestions.

To see this distinction, consider two algorithms that might be deployed in the immigration context. One is optimized to predict the true trial appearance rate for a person detained by ICE. The other aims to optimize uptake by officers, so that the measure of a prediction's quality is whether or not it is accepted by the officer. These two paths may lead to important differences in model performance — and to important legal differences as well. While an algorithm built for accuracy informs and ultimately sways an adjudicator's thinking, it does so via the permissible path of providing a rational summary of available information. By contrast, an algorithm designed to persuade is much more likely to induce automation bias, and conversely is much less defensible from the standpoint of rational government. These factors should both weigh in favor of finding such systems to be "rules." Indeed, part of what is so insidious about algorithms trained primarily to persuade is that they can shift users' preferences.[149] If proven to be scalable, this provides a clear way to manipulate agency adjudicators into silently implementing

[148] Glaze et al., *supra* note 11, at 3.

[149] *See* Micah Carroll, Anca Dragan, Stuart Russell & Dylan Hadfield-Menell, *Estimating and Penalizing Induced Preference Shifts in Recommender Systems*, 162 PROC. MACH. LEARNING RSCH. 2686 (2022), https://proceedings.mlr.press/v162/carroll22a/carroll22a.pdf [https://perma.cc/6CXA-RCG2].

de facto rules. Why go through rulemaking when an algorithm can find a way to make adjudicators prefer your policy en masse?

Another set of problems might arise when these objectives are combined. Conflicts in the optimization objective can leave the receiver (i.e., the adjudicator) worse off.[150] That is because such disagreements may motivate the sender to produce a garbled, rather than informative, signal.[151] The RCA algorithm deployed to generate recidivism predictions for immigrant detainees offers an example of this phenomenon at work. Recall that "resistance by ICE officers [to adopting RCA recommendations] led to high rates of dissent and frequent algorithm edits, where policymakers deleted, added, and reweighted items in the risk assessment tool" to improve compliance.[152] Rather than optimizing for accuracy, algorithm designers opted to garble the signal in an effort to increase the influence of the algorithm.[153] The result left adjudicators using a far less useful algorithm.

F. Human Factors: Time Pressure, Task Complexity, Confidence, and Social Accountability

As we emphasize throughout this Part, additional nonalgorithmic factors will have interaction effects with the other components we discuss here.[154] One might think of at least some of these factors as mere extensions of the elements we have already discussed above. Take time pressure, which we know from studies of aviation and medicine to be highly correlated with the tendency to defer to algorithmic predictions.[155] Time pressure is pervasive in the administrative state, such as when adjudicators are required to complete a certain number of cases in a given time period, or when they receive performance bonuses for output volume. In some sense, time pressure is a way of reducing decision-makers' access to auxiliary information. The less time an adjudicator has, the less free they have to compare a recommendation to alternative evidence from the record. It is thus no surprise that time pressure leads to more reliance on algorithmic tools.

[150] *See id.* at 2687.

[151] *See id.*

[152] Robert Koulish & Ernesto Calvo, *The Human Factor: Algorithms, Dissenters, and Detention in Immigration Enforcement*, 102 SOC. SCI. Q. 1761, 1765 (2021).

[153] *See id.* (noting that "policymakers deleted, added, and reweighted items in the risk assessment tool with the objective of lowering dissent by officers and supervisors" rather than targeting ground-truth risk signals).

[154] *See, e.g.*, De-Arteaga et al., *supra* note 90, at 2.

[155] *See, e.g.*, Nadine B. Sarter & Beth Schroeder, *Supporting Decision Making and Action Selection Under Time Pressure and Uncertainty: The Case of In-Flight Icing*, 43 HUM. FACTORS 573, 574 (2001); Kate Goddard, Abdul Roudsari & Jeremy C. Wyatt, *Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators*, 19 J. AM. MED. INFORMATICS ASS'N 121, 125 (2012).

Time pressure is far from the only example of an external condition that shapes the influence of algorithmic tools. The complexity of adjudicators' tasks under these time pressures,[156] the decision-maker's degree of self-confidence,[157] the amount of training or experience a decision-maker has,[158] and the presence of social pressures or accountability might all play a role in determining how heavily the decision-maker depends on the model.[159] In one recent study, it was found that experienced users were more likely to calibrate their trust in the system to the system's performance, whereas inexperienced users would suffer from overreliance on the system "due to their lack of proper knowledge to detect errors."[160] Another study found that users were more likely to defer to an algorithm when making decisions on "out-of-distribution data" than for "in-distribution data," suggesting that when users are unfamiliar or untrained for a given setting, they will defer to the algorithm, even if it, too, performs worse in that setting.[161] And numerous studies have found that, as humans become accustomed to using an algorithm, they may become inattentive.[162] This is sometimes referred to as automation-induced complacency.[163]

[156] *See id.* at 125; Eric Bogert, Aaron Schecter & Richard T. Watson, *Humans Rely More on Algorithms than Social Influence as a Task Becomes More Difficult*, 11 SCI. REPS. 8028, at 1, 6 (2021) (finding that, "for intellective tasks, humans are more accepting of algorithmic advice relative to the consensus estimates of a crowd").

[157] *See, e.g.*, John D. Lee & Neville Moray, *Trust, Self-Confidence, and Operators' Adaptation to Automation*, 40 INT'L J. HUM.-COMPUT. STUD. 153, 177 (1994).

[158] *See, e.g.*, De-Arteaga et al., *supra* note 90, at 9; Katharina Marten, Tobias Seyfarth, Florian Auer, Edzard Wiener, Andreas Grillhösl, Silvia Obenauer et al., *Computer-Assisted Detection of Pulmonary Nodules: Performance Evaluation of an Expert Knowledge-Based Detection System in Consensus Reading with Experienced and Inexperienced Chest Radiologists*, 14 EUR. RADIOLOGY 1930, 1936 (2004); Kate Goddard, Abdul Roudsari & Jeremy C. Wyatt, *Automation Bias: Empirical Results Assessing Influencing Factors*, 83 INT'L J. MED. INFORMATICS 368, 373 (2014).

[159] *See, e.g.*, Linda J. Skitka, Kathleen L. Mosier, Mark Burdick & Bonnie Rosenblatt, *Automation Bias and Errors: Are Crews Better than Individuals?*, 10 INT'L J. AVIATION PSYCH. 85, 87 (2000).

[160] Mahsan Nourani, Joanie T. King & Eric D. Ragan, *The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems*, 2020 PROC. EIGHTH AAAI CONF. ON HUM. COMPUTATION & CROWDSOURCING 112, 112, https://arxiv.org/pdf/2008.09100.pdf [https://perma.cc/L8LX-C4R7].

[161] *See* Chun-Wei Chiang & Ming Yin, *You'd Better Stop! Understanding Human Reliance on Machine Learning Models Under Covariate Shift*, 2021 PROC. THIRTEENTH ACM WEB SCI. CONF. 120, 120, https://dl.acm.org/doi/abs/10.1145/3447535.3462487 [https://perma.cc/GLQ9-4UAK].

[162] Dietrich Manzey, J. Elin Bahner & Anke-Dorothea Hüper, *Misuse of Automated Decision Aids: Complacency, Automation Bias and the Impact of Training Experience*, 66 INT'L J. HUM-COMPUT. STUD. 688, 689 (2008) ("A typical example involves pilots who rely on the proper function of their autopilot so much that they neglect to monitor and check its function appropriately.").

[163] *Id.* at 688–89.

In another recent study, researchers showed that humans would rather pay an economic cost than bargain with an AI system.[164] This suggests that structuring an algorithm in a way that requires humans to battle with an algorithm's decision-making process might lead them to simply opt out.

These additional factors should all be taken into consideration when designing an algorithmic system and when distinguishing between guidance and rules. And the time pressures are very real in federal agencies. As a recent study of the Board of Veterans' Appeals notes, a veteran's appeal typically takes five years, and the administrative judge handling these appeals "often has no more than an hour to review thousands of pages in the record."[165] These crushing time pressures, which some administrative judges experience as preventing them from deciding cases with "integrity,"[166] might inoculate users from feeling discomfort with surrendering their decisional authority to an algorithm. After all, one might think, administrative judges are already unable to provide due process under the status quo. What's so bad about relying on an expert-crafted algorithm when plan B is to do a rushed job yourself?

Under a "practically binding" test, any algorithm that provides an easy route to a quick decision will likely take over any discretionary decision-making under such pressures. And in such cases, it should likely be treated as a rule, since it is highly likely to effectively replace the exercise of discretion by line adjudicators. That being said, the algorithm would likely be an improvement over the status quo if sufficiently robust and well-tested. A recent meta-analysis found that humans were better decision-makers when provided with algorithmic assistance.[167] So relegating the algorithm to the rulemaking process might result in a net welfare loss if the rulemaking process is lengthy. For this reason, we discuss potential reforms briefly in Part V below.

G. Ex Post Factors

Ex post factors are the set of information that becomes available to courts after an algorithm has been implemented. Recall from Section III.A that courts conducting ex post analysis have looked to factors like

[164] *See* Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand & Ujwal Gadiraju, *For What It's Worth: Humans Overwrite Their Economic Self-Interest To Avoid Bargaining with AI Systems*, 2022 PROC. CHI CONF. ON HUM. FACTORS COMPUTING SYS., at 1, 12, 15, https://dl.acm.org/doi/10.1145/3491102.3517734 [https://perma.cc/D4ET-EBPK].

[165] Ames et al., *supra* note 11, at 4.

[166] *Id.* (citations omitted).

[167] Schemmer et al., *supra* note 122, at 624.

how many exceptions to the policy were granted by the agency,[168] whether significant shifts in officer behavior coincided with algorithmic updates,[169] and whether "affected private parties [were] reasonably led to believe that failure to conform [would] bring adverse consequences."[170] Each of these factors reflects a functional investigation into whether a document has proven binding in fact.

This was precisely the kind of inquiry that Judge Hellerstein undertook in *Velesaca v. Decker*,[171] in which an immigrant detainee alleged that ICE's RCA algorithm should have gone through notice and comment.[172] The core of the plaintiff's challenge focused on the existence of an unstated "No-Release" policy: where the RCA algorithm had once recommended a number of potential outcomes including release, ICE management had allegedly silently changed the algorithm so that it *never* recommended release.[173]

Judge Hellerstein's primary focus was on the degree to which the "No-Release" policy limited officer discretion. If officers had continued to make individualized custody findings and override the RCA algorithm's suggestions, then many APA claims regarding the RCA algorithm itself might "evaporate."[174] But the court accepted statistical evidence to the contrary. Although ICE officials claimed that there was no "No-Release" policy in place,[175] the court instead credited the statistical evidence that the rapid increase in detention rates were coincident with ICE management's change in the algorithm — and therefore

[168] *See, e.g.*, Texas v. United States, 809 F.3d 134, 172 (5th Cir. 2015), *aff'd*, 579 U.S. 547 (2016) (noting, while analyzing factors indicating that DAPA was a substantive rule, that for DACA, a comparative datapoint to DAPA, "5% of the 723,000 applications accepted for evaluation had been denied," and the government did not provide any data on how many of these denials were based on officer discretion).

[169] Velesaca v. Decker, 458 F. Supp. 3d 224, 241–42 (S.D.N.Y. 2020) (noting "[p]er the data, ICE went from releasing (on bond or recognizance) upwards of 30% of alien arrestees to releasing around 2%," and that this was indicative of a "No-Release" policy which should have gone through notice and comment).

[170] Gen. Elec. Co. v. EPA, 290 F.3d 377, 383 (D.C. Cir. 2002); *see also Texas*, 809 F.3d at 173 (noting that "the President had made public statements suggesting that in reviewing applications pursuant to DAPA, DHS officials who 'don't follow the policy' will face 'consequences,' and 'they've got a problem.'").

[171] 458 F. Supp. 3d 224 (S.D.N.Y. 2020).

[172] *See id.* at 239, 242.

[173] *See id.* at 227. It is important to note that while this was not known in *Velesaca*, litigation in a different case revealed that officers were not informed that the RCA algorithm had been updated such that it could not recommend the release of an individual. *Ramirez v. ICE*, 471 F. Supp. 3d 88, 187 (D.D.C. 2020) ("Officers are instructed to run a Risk Classification Assessment tool that, unbeknownst to many officers, never recommends release.").

[174] *Velesaca*, 458 F. Supp. 3d at 239 ("The centrality of this factual issue cannot be overstated. If the No-Release Policy does *not* exist, as Defendants contend, then Plaintiffs' challenges largely evaporate. For example, if ICE officers are in fact making individualized custody findings, then there is no need for the Court to impose an order requiring ICE officers to make such findings; if ICE is not denying release on bond as a matter of general policy, then there has been no *sub silentio* shift in policy in contravention of the APA. And so on.").

[175] *Id.* at 242.

caused by it.[176] In short, the court concluded, "the numbers speak for themselves."[177] The RCA was a rule.

Legally, *Velesaca* offers an evocative example of muscular ex post review at work, and suggests that statistical evidence could play a significant role in judicial review. The core message is that if officers cannot, or do not, exercise discretion, the regime that supplied their true decisional criteria must be subject to notice and comment procedures. Further, just as in the DAPA case, *Velesaca* allows statistical evidence to prove the absence of discretion. Recall that the Fifth Circuit, in analyzing DAPA, had focused on the fact that "5% of the 723,000 applications accepted for evaluation had been denied," although the court acknowledged that it was unclear what share of the denials had been the result of officer discretion.[178] In both cases, statistical evidence concerning deviation from an allegedly uniform policy made identifying the absence of discretion a relatively straightforward proposition.

But even if *Velesaca* stands for the legal proposition that ex post identification of rules is possible, the empirical evidence in *Velesaca* shows just how rarely the stars might align to allow it. The *Velesaca* plaintiffs were able to identify a single, discontinuous, and major change in the RCA algorithm that could be used to test their empirical theory about the "No-Release" policy.[179] Indeed, changes to the RCA algorithm happened regularly: "[P]olicymakers deleted, added, and reweighted items in the risk assessment tool with the objective of lowering dissent by officers and supervisors."[180] Each of these algorithmic updates provided a potential source of variation to undertake a causal study. Of course, a multitude of background factors — like the lag time for officers to perceive the new standards, exogenous changes in immigration flows, and potential deviation from the change — threatened to cloud the analysis of any individual change. But, once again, the *Velesaca* plaintiffs were "lucky" to study a change whose magnitude was large enough to overcome other sources of noise in the data.

Other plaintiffs may not be so lucky. True, algorithms are often quite persuasive, as we argue throughout this paper, and one should expect that to result in major behavioral change when algorithms are implemented.[181] But in the vast majority of cases, the introduction (or modification) of algorithms will coincide with other institutional

[176] *Id.* at 241.

[177] *Id.*

[178] Texas v. United States, 809 F.3d 134, 172 (5th Cir. 2015), *aff'd*, 579 U.S. 547 (2016).

[179] *Velesaca*, 458 F. Supp. 3d at 241.

[180] Koulish & Calvo, *supra* note 152, at 1765.

[181] *See, e.g.*, De-Arteaga et al., *supra* note 90, at 9; Ben Green & Yiling Chen, *Algorithmic Risk Assessments Can Alter Human Decision-Making Processes In High-Stakes Government Contexts*, PROC. ACM ON HUM.-COMPUT. INTERACTION, Oct. 2021, at 1, 18–19; Chiang & Yin, *supra* note 161, at 128.

changes that make it difficult to tell whether changes in outcomes are due to the deployment of the algorithm.[182]

The coincidence of several potential explanations for a change in officers' behavior necessitates resorting to a causal identification strategy of some kind — an analog or alternative to the discontinuity-in-time approach taken in *Velesaca.* But attempts to disentangle potential confounders to identify causal relationships between a treatment and an outcome are notoriously challenging, especially in law.[183] In this case, the goal of a "practically binding" test is to identify whether the introduction of the algorithm led officers to change their behavior to such a degree that they were not exercising their discretion.

Other potential identification strategies could include varying the recommendations that officers get, or only revealing algorithmic decisions to some officers. Then there could be more control in identifying whether officers exercise discretion significantly differently with and without the algorithm — to the point where it is effectively binding. One might also look to whether officers using the algorithm reduce decision times so much that it is not possible that they have truly exercised discretion. For example, one might expect an officer who is rubber-stamping an algorithmic decision to simply click through quickly without looking at any other evidence. On the other hand, an officer truly exercising discretion might look at the algorithmic recommendation and then check some sources of evidence first.

But there are two problems. First, the data required to conduct such an ex post analysis of algorithmic deference is not likely to exist. Agencies may simply not collect the data, or it may not be easy to acquire for would-be plaintiffs to conduct a third-party analysis. And even if the data exists, ex post causal analysis can be tricky, if not impossible, if there is no available causal identification strategy. In the *Velesaca* context, this analysis was possible because the agency recorded enough data for plaintiffs to investigate, but other agencies in the federal government might not do so. And the analysis was convincing because of the large shift in policy. Smaller changes over a longer period of time would be more difficult to discern. Ironically, in this way, algorithms can provide a cover to avoid the rulemaking process by slowly rolling out a policy in a way that is hard to causally identify.

Second, it is unclear where the line is drawn here. What percentage of officers must defer to an algorithm for it to be de facto rulemaking?

[182] *See* Daniel E. Ho & Donald B. Rubin, *Credible Causal Inference for Empirical Legal Studies*, 7 ANN. REV. L. & SOC. SCI. 17, 22 (2011) (arguing that the "credibility of unfoundedness" is "a qualitative judgment that depends crucially on substantive knowledge" that is legal rather than mathematical).

[183] *Id.* at 22, 26; *cf.* Joshua D. Angrist & Jörn-Steffen Pischke, *The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics*, 24 J. ECON. PERSPS. 3, 6 (2010) (discussing how causal strategies are necessary to identify potential effects).

Velesaca involved a completely pervasive culture of deference with virtually uniform adherence to algorithm outputs. Would something less — say, agreement with the algorithm in *most* cases — be sufficient? Fundamentally, we cannot answer these questions. Reliance on ex ante factors, informed by HCI and Bayesian persuasion literature, can help to explain whether a high agreement rate is an indication that an algorithm is or is not practically binding.

V. How the Practically Binding Test Might Motivate Better Algorithmic Practice

Paying attention to the informational design of the systems at the heart of algorithmic governance not only is a reasonable way to approach the doctrinal task of distinguishing rules and guidance, but it may also produce positive incentives for agencies to adopt algorithmic tools that reflect values like transparency and reason-giving. Whatever the drawbacks of notice and comment rulemaking as a paradigm for regulating algorithms in government,[184] adopting a Bayesian persuasion approach to defining discretion is most faithful to the law — and is most likely to push agencies toward designing tools that keep humans in the driver's seat.

In this way, the APA might yet play a constructive role in the development of algorithms in government. Researchers have consistently pointed out that keeping humans in the decision-making loop is important for safety and for accuracy.[185] Much as autopilot on an airplane can be a critical tool for safety if pilots are attentive and well-trained, so too must humans be engaged, informed, and actually exercising their discretion to prevent safety failures. The alternative is overreliance on the output of the algorithm, otherwise known as automation bias. The danger of automation bias is that adjudicators learn to rubber-stamp their algorithmic tools, missing mistakes that humans would catch if they were fully engaged.[186]

A clear example of this is a recent adjudicatory catastrophe in the Netherlands.[187] The Dutch Tax Authority used a machine learning

[184] *See supra* Section II.B for a discussion of the many critiques of notice and comment rulemaking as a tool for regulating algorithms.

[185] *See, e.g.*, De-Arteaga et al., *supra* note 90, at 10 (emphasizing that "providing humans with autonomy to contradict the machine mitigated the effects of miscalculated scores," and arguing that design recommendations "should focus on augmenting the human's ability to identify and correct mistakes").

[186] *See, e.g.*, *id.* at 2 ("Users affected by *automation bias* . . . will follow tool recommendations despite available (but unnoticed or unconsidered) information that would indicate that the recommendation is wrong.").

[187] Rahul Rao, *The Dutch Tax Authority Was Felled by AI — What Comes Next?*, IEEE SPECTRUM (May 9, 2022), https://spectrum.ieee.org/artificial-intelligence-in-government [https://perma.cc/N2M4-MFXT].

algorithm to process childcare benefits applications.[188] When a family uploaded an application to claim government childcare allowance, the algorithm evaluated the claim for signs of fraud, and then humans would review any flagged high-risk claims.[189] Unsurprisingly, the algorithm made a huge number of mistakes, calling fraud where there was none.[190] And though humans were involved, civil servants ended up deferring to the algorithm much of the time.[191] As a result, the tax authority "baselessly ordered thousands of families to pay back their claims, pushing many into onerous debt and destroying lives in the process."[192] The effects of these mistakes were devastating. "Tens of thousands of families — often with lower incomes or belonging to ethnic minorities — were pushed into poverty because of exorbitant debts to the tax agency. Some victims committed suicide. More than a thousand children were taken into foster care."[193]

Avoiding the kind of automation bias that plagued the Dutch Tax Authority is critical to the credible use of algorithms in government. And the APA's focus on the presence of discretion offers one pathway to flex an existing regulatory tool to meet that goal. After all, if agencies are incentivized to avoid burdensome APA review by implementing some of the strategies discussed above — like encouraging disagreement with algorithmic recommendations, designing tools to offer easy access to critical information, and giving adjudicators the time to thoroughly review recommendations — then adjudicators might be less likely to rubber-stamp mistakes the way that Dutch tax examiners did.

That is not the only possible productive effect of APA review. By providing plaintiffs with a cause of action, the APA gives them access to discovery, at least to the limited extent permitted for review of the administrative record. While the scope of that discovery might be limited, it is nonetheless a powerful tool for extracting information from agencies, which may in turn be key to bringing other legal protections to bear.[194] And importantly, it allows independent algorithmic experts

[188] *Id.*

[189] *Id.*

[190] *Id.*

[191] *Id.* ("[T]he algorithm developed a pattern of falsely labeling claims as fraudulent, and harried civil servants rubber-stamped the fraud labels.").

[192] *Id.*

[193] Melissa Heikkilä, *Dutch Scandal Serves as a Warning for Europe over Risks of Using Algorithms*, POLITICO (Mar. 29, 2022, 6:14 PM), https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/ [https://perma.cc/V93M-J73E].

[194] *See supra* Section II.B (discussing the potential utility of APA suits in obtaining information about AI systems).

to weigh in on potential safety issues in algorithmic design in a way that agencies are required to address.[195]

We also note that excessive reliance on algorithmic outputs might also bleed into arbitrary-and-capricious review. Consider, for example, that an agency adopting an algorithm as a rule, or a court treating the algorithm as a de facto rule, may have to justify the algorithmic design. Tools like the RCA algorithm at issue in *Velesaca* (even prior to the "No-Release" edits) might face significant hurdles upon arbitrary-and-capricious review. Conflict between the Obama Administration and ICE workers "led to . . . frequent algorithm edits, where policymakers deleted, added, and reweighted items in the risk assessment tool with the objective of lowering dissent by officers and supervisors."[196] As a result, the algorithm was effectively trying to mimic adjudicators. It was not trained to optimize a legitimate statutory objective, like reducing the no-show rate or crime on release. It is not clear that an algorithm designed to minimize dissent by line adjudicators, but not to optimize factors identified as relevant by statute, would be a legitimate use of the agency's discretion.

Obviously, notice and comment is not all roses. While others have discussed the shortcomings of notice and comment in the context of algorithms, we note three main points: (1) It favors the status quo, preventing the fast iteration so critical to appropriate algorithmic development;[197] (2) it is expensive, and may therefore prevent the kind of continuous revision suited to a new toolkit like administrative algorithms;[198] and (3) it may not require much specificity such that the rulemaking process does not restrict algorithmic use in any meaningful way.[199] Most fundamentally, APA review places the institutional

[195] Experts have recently emphasized the need for independent third-party audits of AI systems by external technical experts. Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg & Daniel Ho, *Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance*, 2022 PROC. AAAI/ACM CONF. ON AI, ETHICS, & SOC'Y 557, 565–66, https://arxiv.org/abs/2206.04737 [https://perma.cc/YT7J-MVR9] ("The literature strongly supports training, standardization, and accreditation for third-party AI auditors."). Rulemaking provides a mechanism for third-party experts to weigh in on an algorithm's design, if described in enough detail. Importantly, the "relevant matter presented" in a comment by technical experts must receive consideration by the agency. 5 U.S.C. § 553(c).

[196] Koulish & Calvo, *supra* note 152, at 1765.

[197] Engstrom & Ho, *supra* note 2, at 821; *see also* MISO Transmission Owners v. Fed. Energy Regul. Comm'n, 45 F.4th 248, 264 (D.C. Cir. 2022) (explaining that, under the APA, an agency is "entitled to change its mind" only if it "provide[s] a 'reasoned explanation' for its decision to disregard 'facts and circumstances that' justified its prior choice." (quoting FCC v. Fox Television Stations, Inc., 556 U.S. 502, 515 (2009))).

[198] Further, the non-adoption of algorithms might mean leaving in place systems whose inefficiency violates rights, including due process rights. *See, e.g.*, Ames et al., *supra* note 11, at 24; Glaze et al., *supra* note 11, at 3.

[199] For example, an agency could simply say that they will use an algorithm to accomplish some task without truly specifying the algorithm's functionality or organizational setting. Of course, comments might reveal deficiencies that would have to be addressed, but there is a level of strategic ambiguity that must be overcome.

strengths and weaknesses of the judiciary, with its frequent attention to individual cases rather than to system-wide goals, at the heart of policy development.

These characteristics, and especially the APA's status-quo bias, are especially worrisome from an algorithmic safety perspective. Algorithmic safety requires continuous iteration to align an algorithm to shifting contexts and new best practices.[200] An agency would be bound to the notice and comment process to make revisions to its algorithm, which could cause harm if an urgent algorithmic failure needs to be fixed, or if regular revisions are needed.[201] "An agency may not, for example, depart from a prior policy *sub silentio* or simply disregard rules that are still on the books."[202] Interestingly, if ICE were to officially announce the RCA algorithm's role in release adjudications as a rule, it is uncertain whether officers would be able to deviate from that process at all after that point. Under the *Accardi* doctrine,[203] ICE would be bound to the rules it has announced for itself.[204]

For the reasons above, rulemaking plays a two-sided role. Despite its potential value for oversight, defending notice and comment as the *optimal* regulatory regime for algorithms would be a heavy burden indeed.[205] We instead emphasize that the specter of rulemaking under the current "practically binding" test may have value in nudging agencies toward known best practices in algorithmic design.

VI. Conclusion: Looking Ahead

To conclude, we briefly summarize our proposed test for courts and emphasize how the highlighted factors in Part IV might provide a constructive impetus for agencies to adopt safer algorithms in order to avoid the strictures of notice and comment rulemaking. We also briefly discuss legislative reforms that would improve the rulemaking process.

[200] *See, e.g.*, Engstrom & Ho, *supra* note 2, at 821; Peter Henderson, Ben Chugg, Brandon Anderson & Daniel E. Ho, *Beyond Ads: Sequential Decision-Making Algorithms in Law and Public Policy*, 2022 Proc. ACM Symp. on Comput. Sci. & L. 87, 96–97, https://dl.acm.org/doi/10.1145/3511265.3550439 [https://perma.cc/5LYQ-38CC].

[201] We note, though, that there are of course emergency procedures that act as safeguards such as the "good cause exception." *See, e.g.*, 5 U.S.C. § 553(b)(B), (d); Kevin Hartnett Jr., *An Approach to Improving Judicial Review of the APA's "Good Cause" Exception to Notice-and-Comment Rulemaking*, 68 Buff. L. Rev. 1561 (2020) (discussing the applicability of the good cause exception to different emergency situations and how courts tend to review such situations).

[202] FCC v. Fox Television Stations, Inc., 556 U.S. 502, 515 (2009).

[203] United States *ex rel.* Accardi v. Shaughnessy, 347 U.S. 260, 274 (1954).

[204] *See* Steenholdt v. Fed. Aviation Admin., 314 F.3d 633, 639 (D.C. Cir. 2003) ("The *Accardi* doctrine requires federal agencies to follow their own rules, even gratuitous procedural rules that limit otherwise discretionary actions.").

[205] *See* Engstrom & Ho, *supra* note 2, at 839, for an extensive critique along these lines.

A. Approaching Algorithms from a Judicial Standpoint

The factors we have described here should directly inform judicial decisions about whether an algorithm is practically binding.

First, courts should pay close attention to the design of the algorithm. The most important factor, in our view, is the extent to which an algorithm's interface maps onto the ultimate decision that an adjudicator has to make. For example, if an algorithm implemented by ICE to decide on bail gives hearing officers a recommendation on the amount of the recommended bond, it will be all too easy for the officers to wholly accept that suggestion. On the other hand, if an algorithm merely recommends evidence for an officer to examine, legal frameworks to apply, or even questions to ask, then the officer will not be in a position to plug the output of the algorithm into a decision document. In short, the more cognitive processing needed to get from the algorithm to the final decision, the further an algorithm is from a rule.

Second, the immediate context of the algorithm's recommendation matters, too. If a software tool pairs the algorithm's recommendation with contextual information — like legal references, excerpts from the record, or other evidence that would be relevant to the decision — that would reduce the costs of thinking critically about the algorithm's recommendation. An unadorned recommendation (e.g., a screen with the word "BAIL" written on it) would do the opposite, requiring the decision-maker to start from scratch in constructing an independent view. Making it easy for decision-makers to form their own conclusions from the underlying data is a key part of preserving discretion.

Finally, the likelihood of automation bias significantly depends on the real-world processes surrounding the algorithm. The press of an officer's caseload is a major factor. So are the costs and benefits of adherence to the algorithm's recommendations. If agencies punish mistakes more harshly when they conform to an algorithm's recommendations, then officers may be more hesitant to defer. More likely, however, is that humans will be punished for disagreeing with statistical recommendations. That worrisome arrangement may quickly turn "recommendations" into rulings.

To sum up, the overall question courts should ask is how hard it is for decision-makers to dissent from the algorithm's view. If it would take enormous self-discipline, unrewarded effort, and professional risk to add an independent view to an algorithm's recommendation, notice and comment is more appropriate.

B. The View from Agencies

These same factors should inform how agencies implement best practices in algorithmic safety.

The most difficult internal battle may be overcoming the desire to prove to internal stakeholders that algorithms make processing faster or easier. It is true, of course, that well-designed algorithms should vastly increase the efficiency of decision-making processes. But a procurement process designed around maximizing the speed gains from an algorithmic support tool is a recipe for complacency and rubber-stamping. By contrast, our approach suggests that intentionally increasing cognitive frictions in transforming an algorithmic output into a final decision — for instance, by focusing on recommended sources rather than recommended results — may be necessary to avoid reducing discretion to a nullity.

Furthermore, line officers ought to be involved in the design phase to ensure maximum input on the presentation of contrary evidence, auxiliary information, and other surrounding information to reduce reliance on algorithmic outputs.

Most importantly, agencies must carefully design policies to encourage dissent from algorithms. As a cultural matter, dissent from algorithmic recommendations should be prized and encouraged. The best way to do that is to ensure that costs be uniform or align in favor of discretion. For example, many agencies assign officers credits corresponding to the number of decisions they issue. Agencies could assign more credits to officers whose patterns of decision-making demonstrates discretion. And if written reasoning is required, it should not be only when officers deviate from the algorithm, lest this requirement incentivize automation bias.

Like Engstrom and Ho, we call for agencies to aggressively collect data to empirically verify that officers actively exercise discretion, ideally through randomized rollouts of new tools.

In an ideal world, agencies would be less averse to rulemaking because it would be a less taxing process, especially for algorithms. Making that change would require legislative reform. While we acknowledge the challenges of that path forward, we join the chorus of voices deeply concerned that the rulemaking process is too rigid to deal well with algorithms. While the specter of rulemaking might encourage agencies to follow best practices in some regards, the focus of rulemaking on stasis and consistency might introduce perverse incentives as well. Reforms should include faster turnaround times with less onerous procedure, lowered costs, and specific requirements on transparency and evaluation mechanisms. Such a revised algorithmic rulemaking process, when combined with the functionalist approach we describe here, would ensure that either pathway conforms to safe algorithm deployments. Either there is an attentive and engaged human in the loop, or the model has been vetted by a thorough, well-defined process.