NOTES

MACHINE RULEMAKING: ARBITRARY AND CAPRICIOUS REVIEW IN THE AGE OF AI

In the past few years, hype surrounding artificial intelligence (AI) and machine learning (ML) has reached a fever pitch. From newspapers to academic journals, it's nearly impossible to avoid predictions of either the technology's boundless potential[1] or its catastrophic effects.[2] Moreover, use of AI/ML is no longer limited to tech companies or the private sector: In 2020, a report prepared for the Administrative Conference of the United States documented "157 [AI/ML] use cases across 64 agencies."[3] And just this year, the Department of Government Efficiency released a custom AI chatbot to 1,500 federal employees.[4]

In some sense, federal agency adoption of ML is unsurprising. The technical expertise and flexibility of agencies are commonly touted as some of their key strengths.[5] In this vein, experimentation with ML might simply be proof that agencies are living up to their full potential. Indeed, from automating rote tasks to analyzing vast, unstructured data sets with greater accuracy, ML has the "potential to reduce the cost of core governance functions, improve the quality of decisions," and generally "mak[e] government performance more efficient and effective."[6] The U.S. Treasury, for example, used ML to recover $1 billion in check fraud over the course of just one year.[7] However, new technologies also

---

[1] *See, e.g.*, Cade Metz, *Google Unveils A.I. for Predicting Behavior of Human Molecules*, N.Y. TIMES (May 8, 2024), https://www.nytimes.com/2024/05/08/technology/google-ai-molecules-alphafold3.html [https://perma.cc/N598-FVAT]; Fabio Boniolo et al., *Artificial Intelligence in Early Drug Discovery Enabling Precision Medicine*, 16 EXPERT OP. ON DRUG DISCOVERY 991, 992 (2021); Herbert B. Dixon Jr., *Artificial Intelligence: Benefits and Unknown Risks*, JUDGES' J., Winter 2021, at 41, 42 ("AI technology has the potential to transform forensic science in criminal investigations significantly.").

[2] *See, e.g.*, Aaron Gregg et al., *AI Poses "Risk of Extinction" on Par with Nukes, Tech Leaders Say*, WASH. POST (May 30, 2023), https://www.washingtonpost.com/business/2023/05/30/ai-poses-risk-extinction-industry-leaders-warn [https://perma.cc/56EW-PZYC]; Eden Sarid & Omri Ben-Zvi, *Machine Learning and the Re-Enchantment of the Administrative State*, 87 MOD. L. REV. 371, 391 (2024) ("[W]hen . . . an algorithm's prediction is used by an administrative body, something is lost.").

[3] DAVID FREEMAN ENGSTROM ET AL., GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 16 (2020).

[4] Makena Kelly & Zoë Schiffer, *DOGE Has Deployed Its GSAi Custom Chatbot for 1,500 Federal Workers*, WIRED (Mar. 7, 2025, 6:22 PM), https://www.wired.com/story/gsai-chatbot-1500-federal-workers [https://perma.cc/38JZ-9FGG].

[5] *See* JOHN F. MANNING & MATTHEW C. STEPHENSON, LEGISLATION AND REGULATION: CASES AND MATERIALS 523–24 (4th ed. 2021).

[6] ENGSTROM ET AL., *supra* note 3, at 6.

[7] Press Release, U.S. Dep't of Treasury, Treasury Announces Enhanced Fraud Detection Processes, Including Machine Learning AI, Prevented and Recovered Over $4 Billion in Fiscal Year 2024 (Oct. 17, 2024), https://home.treasury.gov/news/press-releases/jy2650 [https://perma.cc/C3YT-KZZX].

pose new risks. While ML models have demonstrated impressive success,[8] they have also faced notable issues, from reinforcing harmful biases to suddenly becoming inaccurate.[9]

The nature of these risks might differ when ML is used in the public sector as opposed to the private. On the one hand, the potential harm may be intensified: If Facebook's ML model fails, someone might be served an irrelevant ad; if the Social Security Administration's (SSA) ML model fails, someone may be erroneously denied critical benefits.[10] On the other hand, there may also be greater safeguards to prevent such harm, given that agencies are far more regulated than private tech corporations.[11] Agencies must operate within the confines of their organic statutes, and many are subject to the default procedural requirements of the Administrative Procedure Act[12] (APA). But this raises a natural question: Is existing administrative law well positioned to encourage beneficial use of ML while minimizing its potential harm? While courts have long engaged in judicial review of highly technical agency decisions,[13] ML models differ in some unique ways from traditional technology. ML models evolve over time as they learn,[14] and the black box nature of many ML models means that even their designers can't explain why they make the decisions they do.[15] Together, these qualities may make it more difficult (1) for agencies to comply with certain procedural requirements when using ML and (2) for courts to effectively review such use.

---

[8] *See* Bernard Marr, *27 Incredible Examples of AI and Machine Learning in Practice*, FORBES (Apr. 30, 2018, 12:28 AM), https://www.forbes.com/sites/bernardmarr/2018/04/30/27-incredible-examples-of-ai-and-machine-learning-in-practice [https://perma.cc/HT8B-KWXD].

[9] Thor Olavsrud, *12 Famous AI Disasters*, CIO (Oct. 2, 2024), https://www.cio.com/article/190888/5-famous-analytics-and-ai-disasters.html [https://perma.cc/FQ9G-JZKV].

[10] *See, e.g.*, Alex Ebert, *Triple Payouts Approved for Jobless Claims Stripped by Faulty AI*, BLOOMBERG L. (Jan. 29, 2024, 1:07 PM), https://www.bloomberglaw.com/bloomberglawnews/daily-labor-report/X4CIS834000000 [https://perma.cc/AA6Z-P938]. This is not to suggest that private use of AI/ML has not led to serious, pernicious effects in aggregate, but rather that individual ML decisions may have a greater negative impact in the public sector.

[11] *See* Adam Conner et al., *Congress Must Take More Steps on Technology Regulation Before It Is Too Late*, CTR. FOR AM. PROGRESS (May 13, 2024), https://www.americanprogress.org/article/congress-must-take-more-steps-on-technology-regulation-before-it-is-too-late [https://perma.cc/MA97-CBNE].

[12] 5 U.S.C. §§ 551–559, 701–706.

[13] *See, e.g.*, Balt. Gas & Elec. Co. v. Nat. Res. Def. Council, Inc., 462 U.S. 87, 89–90 (1983) (reviewing agency's prediction of nuclear waste leakage from long-term storage); Marsh v. Or. Nat. Res. Council, 490 U.S. 360, 368 (1989) (reviewing agency's analysis of the environmental consequences of constructing a dam and its decision not "to include a 'worst case analysis'").

[14] Judah Phillips, *How Machine Learning Models Improve over Time*, SQUARK (Jan. 14, 2022), https://squarkai.com/how-machine-learning-models-improve-over-time [https://perma.cc/US53-5T9S].

[15] David Freeman Engstrom & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 YALE J. ON REGUL. 800, 824 (2020) ("[E]ven their own engineers cannot necessarily understand how the most advanced models arrived at a given result.").

This Note considers how agency use of ML,[16] specifically in rulemaking, may create novel issues for the APA's § 706(2)(A) arbitrary and capricious review.[17] Given the myriad ways an agency might use ML, this Note proposes three hypothetical examples of ML rulemaking and applies arbitrary and capricious review to each. A few conclusions result from this analysis: First, an agency using ML in informal rulemaking can likely formally satisfy § 706(2)(A) review by providing sufficient "second-order" data about the design, training, and testing of the model. Second, while ML rulemaking may *technically* comply with the APA, this compliance will sometimes fail to achieve the APA's underlying goals of transparency, accountability, and rationality. Third, these limitations suggest that ML rulemaking may be better regulated by shifting focus from ex ante to ex post requirements and encouraging public scrutiny of ML models even after a rule is finalized. Finally, the Note as a whole demonstrates the limitations of making broad generalizations about ML. Even within rulemaking, the precise nature of *how* ML is used will greatly affect the legal implications.

The Note proceeds as follows. Part I provides background by highlighting ways in which ML models differ from traditional technology, summarizing relevant legal scholarship, and introducing three hypothetical case studies of ML rulemaking. Part II provides a detailed analysis of how § 706(2)(A)'s requirements would apply to the case studies, considering the extent to which the agencies could comply with both the letter and spirit of the law. Finally, Part III explores why a greater focus on ex post requirements might better achieve the underlying goals of the APA and how such requirements might be implemented.

I. BACKGROUND

A. Intro to ML

ML models are a specific type of computer algorithm that are particularly effective at analyzing large data sets and extracting patterns and insights.[18] Some ML models make predictions,[19] others classify and

---

[16] This Note focuses on ML models, a popular subset of the "broad field" of AI. *Artificial Intelligence (AI) vs. Machine Learning (ML)*, GOOGLE CLOUD, https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning [https://perma.cc/EZ8N-Q2ZS].

[17] 5 U.S.C. § 706(2)(A).

[18] *What Is Machine Learning (ML)?*, U.C. BERKELEY SCH. OF INFO. (June 26, 2020), https://ischoolonline.berkeley.edu/blog/what-is-machine-learning [https://perma.cc/3AGN-6ZAU].

[19] *What Is Predictive Analytics?*, IBM, https://www.ibm.com/topics/predictive-analytics [https://perma.cc/C26X-GPAF].

categorize data,[20] and still others generate original content.[21] At this level of generality, ML models may sound very similar to traditional algorithms, which also accomplish similar tasks.[22] However, the *way* ML models analyze data and extract insights is quite different. Rather than providing a general overview of ML, this subsection highlights just a few of these key distinctions, which will be important to the subsequent legal analysis.

First, ML models are often described as being "inscrutable"[23] in a way that traditional algorithms are not.[24] This is because even the designers of the system cannot conceptually explain, in any given case, how individual inputs to a model correspond to or affect individual outputs or predictions.[25] On the other hand, some aspects of ML models can be described.[26] For example, engineers can describe virtually every *design* decision they made in building a model, including: the precise kinds of data that the model was trained on, the type of prediction the model was designed to make, how the model functions in the broader technology workflow,[27] and error rates from any validation testing that was conducted.[28] They could even provide a description of the machine's strategy for "learning" over time.[29]

Second, ML models often make predictions that are unintuitive or even seemingly incomprehensible to humans.[30] ML models are often trained on an enormous amount of raw data, allowing them to pick out strange and complex correlations that engineers would never have

---

[20] Badreesh Shetty, *5 Classification Algorithms for Machine Learning*, BUILT IN (Apr. 12, 2023), https://builtin.com/data-science/supervised-machine-learning-classification [https://perma.cc/8GEP-3M6P].

[21] *What Is Generative AI?*, MCKINSEY & CO. (Apr. 2, 2024), https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai [https://perma.cc/9C4V-N4CJ]. ChatGPT is an example. *Id.*

[22] Traditional algorithms can also, for example, make predictions and analyze data. *See, e.g.*, *Linear Regression*, YALE DEP'T OF STAT., http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm [https://perma.cc/R5C7-GQBG].

[23] *See* Engstrom & Ho, *supra* note 15, at 824 ("[A]dvanced machine learning outputs are *inscrutable* . . . ."); Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 COLUM. L. REV. 1851, 1851 (2019).

[24] Traditional inscrutable ML models are the focus of this Note, but it is worth noting that there are active efforts to develop models that are explainable or interpretable. *See* Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 50–55 (2019); Lars Hulstaert, *Black-Box vs. White-Box Models*, MEDIUM: TDS ARCHIVE (Mar. 14, 2019), https://medium.com/towards-data-science/machine-learning-interpretability-techniques-662c723454f3 [https://perma.cc/S9MM-69T9] (describing ML techniques that "enable interpretability").

[25] *See* Engstrom & Ho, *supra* note 15, at 824.

[26] Strandburg, *supra* note 23, at 1872–73.

[27] *See id.*; Coglianese & Lehr, *supra* note 24, at 38–39.

[28] Strandburg, *supra* note 23, at 1873; *see* Coglianese & Lehr, *supra* note 24, at 48–49.

[29] *See* Kizito Nyuytiymbiy, *Parameters and Hyperparameters in Machine Learning and Deep Learning*, TOWARDS DATA SCI. (Dec. 30, 2020), https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac [https://perma.cc/485V-WCRL] (defining "hyperparameters," which "control the learning process" and, in turn, affect how the model will run).

[30] *See* Strandburg, *supra* note 23, at 1852; Coglianese & Lehr, *supra* note 24, at 17.

anticipated.[31] Indeed, these convoluted connections are part of the reason why ML models "can outperform traditional statistical techniques" in the first place.[32]

Third, ML models differ from traditional algorithms in the way that they evolve over time. Traditional algorithms are deterministic: Whether it has been one day or one thousand days, the algorithm will make the same prediction as long as it is fed the same input data.[33] In contrast, an ML model can dynamically learn and adjust its approach as it is used.[34] Again, this is both a source of strength and an area of concern. On the one hand, the ML model can adapt to changing conditions, enabling high-quality predictions over time.[35] On the other hand, even if a model starts out strong, there is no guarantee it will continue to perform well; it might rapidly degenerate if it is poorly designed or if the training data ends up being unrepresentative of the real world.[36]

B. Existing Scholarship

Over the past decade, increasing adoption of AI/ML technology has led to an explosion of legal scholarship on the topic. Much of this work has taken a constitutional lens, considering potential procedural due process, equal protection, and Fourth Amendment challenges.[37] More recently, a growing number of articles have focused on how administrative law applies to government use of ML. This work typically adopts a high level of abstraction — focusing on ML as a general tool, rather than on its specific applications.[38] Substantively, scholarship tends to

---

[31] *See* Strandburg, *supra* note 23, at 1852; Coglianese & Lehr, *supra* note 24, at 17 (describing the "complex inter-variable relationships" that can be uncovered by ML models); Finale Doshi-Velez & Mason Kortz, *Accountability of AI Under the Law: The Role of Explanation* 8 (Berkman Klein Ctr. for Internet & Soc'y, Working Paper, 2017), https://dash.harvard.edu/server/api/core/bitstreams/73120037e-92c5-6bd4-e053-0100007fdf3b/content [https://perma.cc/T9HK-YWZS].

[32] Coglianese & Lehr, *supra* note 24, at 16.

[33] *See AI vs Traditional Algorithms: When to Use AI Models over Classical Methods*, ZIGNUTS TECHNOLAB (Nov. 27, 2024), https://www.zignuts.com/blog/ai-vs-traditional-algorithms [https://perma.cc/Q25G-9VT6].

[34] *See id.*

[35] *See id.*

[36] *See, e.g.*, Olavsrud, *supra* note 9.

[37] *See, e.g.*, Aziz Z. Huq, *Constitutional Rights in the Machine-Learning State*, 105 CORNELL L. REV. 1875, 1879 (2020) (analyzing relationship between ML and "due process, equality, and privacy"); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 95 (2014); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 8 (2014); Cary Coglianese & David Lehr, *Improving the Administrative State with Machine Learning*, ADMIN. & REGUL. L. NEWS, Summer 2017, at 7, 8–9 (considering potential effect of ML rulemaking or adjudication on due process and nondiscrimination principles); Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871, 879 (2016).

[38] *See, e.g.*, Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1177 (2017); Coglianese & Lehr, *supra* note 24, at 5; Sarid & Ben-Zvi, *supra* note 2, at 372; Strandburg, *supra* note 23, at 1858. *But see* Engstrom & Ho, *supra* note 15, at 805 (analyzing actual use of ML by the SSA and SEC).

fall into one of three categories: Some scholars argue that the inscrutable nature of ML creates significant challenges for meeting existing administrative law requirements and erodes agency accountability and transparency.[39] Others argue that concerns about ML inscrutability have been somewhat overblown and that agency use of ML should pose no fundamentally insurmountable issues.[40] Finally, a third group adopts a middle ground, suggesting that agency use of ML may often satisfy existing legal requirements, but that changes to our legal frameworks may be needed to properly regulate ML.[41] This Note seeks to make a small contribution by (1) focusing specifically on rulemaking and (2) considering how § 706(2)(A) review could apply to *specific* hypothetical case studies, thus complicating the more monolithic discussions of ML.

C. Case Studies

This Part introduces three examples of how federal agencies might use ML in informal rulemaking. While inspired by existing data sets and actual use of ML by federal agencies,[42] these cases are ultimately hypothetical, pushing beyond existing use of ML to enable more forward-looking analyses in Part IV.

1. The SEC Rule. — The SEC collects a vast amount of information from the corporations it regulates.[43] This data is provided via extensive reports that present "general business information, risk factors, financial data,"[44] and more, including written descriptions and charts[45] that can be difficult for traditional algorithms to process.[46] Suppose the SEC develops an ML model that processes all such data and then flags corporate filers at a high risk of violating SEC regulations. Applying the model to its existing data, the agency discovers a disproportionate

---

[39] *See, e.g.*, Sarid & Ben-Zvi, *supra* note 2, at 391, 397.

[40] *See, e.g.*, Coglianese & Lehr, *supra* note 24, at 6, 39, 43 ("Machine-learning algorithms' irreducible inscrutability should not form a bar to their use by government officials . . . ." *Id.* at 39.); Coglianese & Lehr, *supra* note 38, at 1205–13 (arguing that ML will not pose an unsurmountable bar to transparency in agency decisionmaking).

[41] *See, e.g.*, Engstrom & Ho, *supra* note 15, at 804, 836 (arguing that existing APA requirements are unlikely to successfully constrain or guide agency use of ML, so a new "accountability structure," *id.* at 804, is needed); Strandburg, *supra* note 23, at 1880–83 (acknowledging unique APA challenges in ML rulemaking, but suggesting that providing descriptive information on the model's components, training data, and validation results could be enough to overcome these challenges); Cameron Averill, *Algorithmic Reason-Giving, Arbitrary and Capricious Review, and the Need for a Clear Normative Baseline*, 93 U. CIN. L. REV. 40, 41 (2024) ("[A]t least some algorithms should pass judicial muster [under the APA].").

[42] These case studies are inspired by actual applications of AI in federal agencies, as documented in a 2020 report commissioned by the Administrative Conference of the United States. *See generally* ENGSTROM ET AL., *supra* note 3.

[43] *Id.* at 23.

[44] *Id.*

[45] *See, e.g.*, Walmart Inc., Annual Report (Form 10-K), at 30 (Mar. 15, 2024).

[46] Alexandra Jonker & Alice Gomstyn, *Structured vs. Unstructured Data: What's the Difference?*, IBM (Feb. 7, 2025), https://www.ibm.com/think/topics/structured-vs-unstructured-data [https://perma.cc/PQ6F-MPJA].

number of "high risk" flags for filers engaging in short-term loaning of securities. Historical enforcement data and the SEC's existing understanding of these transactions affirm the model's finding. The SEC promulgates a rule requiring corporate filers engaging in such lending to provide additional disclosures to the SEC about their activities.[47]

2. The FDA Rule. — After approving a drug, the FDA continues to collect "adverse event" data from drug manufacturers, "[p]atients, caregivers, and healthcare professionals" to monitor how the drug is faring when used by actual patients.[48] These reports can be time-consuming to parse,[49] since they include unstructured, descriptive narratives of the effects on the patient.[50] Suppose the FDA uses these reports to train an ML model that can predict the risk of serious adverse effects for a given drug. The FDA runs the model on a variety of hypothetical drugs and discovers that those that (1) are orally administered, (2) have a specific type of chemical structure, and (3) contain certain ingredients are extremely likely to cause serious adverse effects in children. It is unclear to the FDA why this combination of features would pose such a high risk, and the FDA is unable to support this finding via historical data. However, previous benchmarking of the ML model has shown it to be very successful at predicting adverse effects. Consequently, the FDA promulgates a rule requiring manufacturers to meet additional testing and disclosure requirements for any drug with all three features.

3. The SSA Rule. — The SSA receives millions of disability claims each year and suffers from a "significant backlog of claims."[51] Hundreds of thousands of claims advance to hearings, which may require claimants to wait anywhere "from a few months to more than two years."[52] Suppose the SSA trains a model that can predict the likelihood of a claim being accepted. The model is trained on historical data provided by claimants as well as the ultimate determination made for each claim by the human decisionmaker. In benchmarking tests, the model was just as accurate as human decisionmakers in cases it deemed "highly likely" or "highly unlikely" to be approved.

To reduce its backlog, the SSA promulgates a rule incorporating its model into the procedure for processing claims. The rule states that every claimant seeking disability payments will have their claim information submitted to the model. If the model predicts an extremely high likelihood of approval, the model approves the claim. If the model predicts a very low likelihood of approval, the model "slow tracks" the

---

[47] This is loosely inspired by both the SEC's Corporate Issuer Risk Assessment tool, ENGSTROM ET AL., *supra* note 3, at 23, and a recent rule promulgated by the SEC related to increasing transparency in securities lending, 17 C.F.R. § 240.10c-1a (2024).

[48] ENGSTROM ET AL., *supra* note 3, at 55.

[49] *See id.*

[50] *See id.*

[51] *Id.* at 38.

[52] *Id.*

claim, prioritizing other cases ahead of it. The SSA predicts that doing so will significantly accelerate its overall distribution of benefits.

II. Legal Analysis

While federal agencies are subject to a variety of procedural requirements, the APA looms large by establishing a baseline.[53] Assuming an agency has the statutory authority to promulgate informal rules, the APA imposes two key requirements: First, an agency must follow the procedural steps outlined in § 553,[54] including (1) giving notice to the public of the proposed rule,[55] (2) providing an opportunity for the public to respond to the proposed rule,[56] and finally (3) providing a statement justifying the final rule that addresses significant public comments.[57] Second, § 706(2)(A) requires courts to "set aside" any agency action, including rulemaking, that is "arbitrary" or "capricious."[58]

This Part explores specifically how § 706(2)(A) applies to ML rulemaking. While some scholars have argued that agency use of ML should pose little or no novel issue within the framework of arbitrary and capricious review,[59] this Part argues that it *may*. In particular, even though agencies that use ML may be able to comply with § 706(2)(A) in a narrow, formalistic manner, such compliance could fail to achieve the APA's underlying goals of rationality and transparency.[60]

At the outset, however, it is possible to address the SEC rule. Recall that the SEC rule can be described and justified without ML, as the agency itself was able to understand and confirm the pattern picked out by the model. Consequently, APA requirements may be applied no differently than they would be to a conventional rule. This does not imply, however, that the SEC case study is insignificant. Instead, it is an important reminder that ML rulemaking need not always pose novel legal challenges. In these cases, agencies can and should explore the benefits of ML.[61] Moreover, while this Part focuses on the ways that ML can

---

[53] Manning & Stephenson, *supra* note 5, at 822.

[54] 5 U.S.C. § 553.

[55] *Id.* § 553(b).

[56] *Id.* § 553(c).

[57] *See id.*; United States v. N.S. Food Prods. Corp., 568 F.2d 240, 252 (2d Cir. 1977).

[58] 5 U.S.C. § 706(2)(A).

[59] *See, e.g.*, Coglianese & Lehr, *supra* note 24, at 43–47; Averill, *supra* note 41, at 96; *cf.* Engstrom & Ho, *supra* note 15, at 828 ("[E]x post judicial review of agency action using AI is unlikely to yield systematic scrutiny.").

[60] Due to space constraints, this Note does not consider how § 553 requirements apply to ML rulemaking. However, there are a variety of similar, interesting issues that can be explored (for example, whether the notice-and-comment procedures are equally effective as applied to ML rulemaking, given that the commenting process may become focused on technical, computer science concepts rather than the substantive topic being regulated).

[61] Indeed, when used effectively, ML can lead to significant benefits. *See* Coglianese & Lehr, *supra* note 24, at 16 (describing how ML models "can outperform traditional statistical techniques and . . . surpass the abilities of human decisionmakers").

create unique legal pressures within the APA, the purpose in doing so is not to universally condemn ML — it is far too broad and flexible a tool.[62] Rather, the goal is to explore the extent to which the APA succeeds and fails in promoting successful ML rulemaking by agencies.

This section begins with a general overview of how courts apply the arbitrary and capricious standard to agency promulgation of an informal rule. Next, it reviews why some scholars believe that ML rulemaking can satisfy this standard. Finally, it problematizes these arguments through the lens of the FDA and SSA rules.

A. Doctrinal Overview

Under § 706(2)(A) of the APA, courts are required to "hold unlawful and set aside agency action, findings, and conclusions" that are "arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law."[63] At a high level, courts have interpreted this as a requirement that agencies engage in "reasoned decisionmaking" when rulemaking or adjudicating.[64] Though courts should not "substitute [their] judgment for that of the agency,"[65] they must review an agency's explanation for taking action and ensure there is a "rational connection between the facts found and the choice made."[66] In this sense, § 706(2)(A) implicitly requires a certain level of transparency from agencies so that courts may effectively review their decisions.[67] More specifically, courts should ensure that the agency has given "consideration of the relevant factors" and has not made "a clear error of judgment."[68] An agency action will also "[n]ormally . . . be arbitrary and capricious" if it "relie[s] on factors which Congress has not intended [the agency] to consider" or "entirely fail[s] to consider an important aspect of the problem."[69]

While often referred to as "hard look" review, this standard is fairly deferential in practice.[70] This is especially so when reviewing

---

[62] This would be as absurd as universally condemning all software engineering. Such an adaptable tool can be implemented in both positive and negative ways. *See supra* notes 1–2 and accompanying text.

[63] 5 U.S.C. § 706(2)(A).

[64] Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 52 (1983); *see also id.* at 43; KRISTIN E. HICKMAN & RICHARD J. PIERCE, JR., ADMINISTRATIVE LAW TREATISE § 11.1 (7th ed. Supp. 2024).

[65] *State Farm*, 463 U.S. at 43.

[66] *Id.* (quoting Burlington Truck Lines, Inc. v. United States, 371 U.S. 156, 168 (1962)).

[67] Indeed, courts have specifically interpreted APA notice-and-comment requirements in light of § 706(2)(A)'s prohibition on arbitrary and capricious action. *See, e.g.*, United States v. N.S. Food Prods. Corp., 568 F.2d 240, 252 (2d Cir. 1977).

[68] *State Farm*, 463 U.S. at 43 (quoting Bowman Transp., Inc. v. Ark.-Best Freight Sys., Inc., 419 U.S. 281, 285 (1974)).

[69] *Id.*

[70] *See* ADRIAN VERMEULE, LAW'S ABNEGATION 33–34 (2016) (describing a "counter-canon of deference," which has developed in spite of significant focus on "hard look review" by lawyers

particularly technical or complex agency determinations.[71] In *Baltimore Gas & Electric Co. v. Natural Resources Defense Council, Inc.*,[72] the Supreme Court stated that when reviewing "predictions, within [an agency's] area of special expertise, at the frontiers of science[,] . . . a reviewing court must generally be at its most deferential."[73] Two reasons may justify this deference. First, when making technical, complex decisions, agencies rely on their deep subject matter expertise, so a generalist court should be especially wary of overstepping and "substitut[ing] its judgment."[74] Second, when making these types of decisions "at the frontiers of science," agencies are often operating under genuine uncertainty.[75] Under these circumstances, courts should be careful not to impose a "pathological" demand for "first-order reason[s]," as it may be impossible for the agency to provide them.[76] In *Marsh v. Oregon Natural Resources Council*,[77] for example, the Court acknowledged that "[w]hen specialists express *conflicting* views, an agency must have discretion to rely on the reasonable opinions of its own qualified experts."[78]

B. Why ML Rulemaking Can Comply

Some scholars argue that ML rulemaking should have no theoretical issue satisfying arbitrary and capricious review.[79] First, because ML exists at the very edges of modern technology, the special deference that courts have historically given to agencies when operating at the frontiers of science should certainly extend to ML rulemaking.[80] Second, agencies can provide plenty of second-order information to survive arbitrary and capricious review. Training and design data can demonstrate that the model was created in a thoughtful, rational manner and that it considers all relevant factors.[81] Agencies can also explain any broader "systemic reasons"[82] they have for adopting ML and present validation testing data to defend the model's accuracy or efficiency.[83] This should be more

---

and academics, *id.* at 33 (emphasis omitted)); Coglianese & Lehr, *supra* note 24, at 28 ("Judges by and large do not hold agencies to extremely high standards of rationality under the arbitrary and capricious standard.").

71 Coglianese & Lehr, *supra* note 24, at 28.

72 462 U.S. 87 (1983).

73 *Id.* at 103 (citing Indus. Union Dep't v. Am. Petroleum Inst., 448 U.S. 607, 656 (1980) (plurality opinion); *id.* at 705–06 (Marshall, J., dissenting)).

74 *State Farm*, 463 U.S. at 43.

75 *Balt. Gas*, 462 U.S. at 103; *see, e.g.*, *id.* at 99.

76 VERMEULE, *supra* note 70, at 129.

77 490 U.S. 360 (1989).

78 *Id.* at 378 (emphasis added).

79 *See, e.g.*, Coglianese & Lehr, *supra* note 24, at 47; Averill, *supra* note 41, at 96; *cf.* Engstrom & Ho, *supra* note 15, at 828, 836 (noting the shortcomings of existing ex post review of ML-based agency action).

80 *See* Coglianese & Lehr, *supra* note 24, at 28.

81 *See id.* at 43–44.

82 Averill, *supra* note 41, at 96.

83 *See supra* notes 26–29 and accompanying text.

than enough to satisfy the requirement that there be a rational connection between data and rule. Finally, to the extent that the inscrutability of ML prevents a reviewing court from understanding how a decision was or will be made, these scholars argue that the APA "has never required" the ability to "peer into the minds of government administrators."[84] This argument may seem quite compelling in the abstract. However, the next section considers potential issues raised by the case studies.

C. Challenges with ML Rulemaking

1. The Limits of Second-Order Review. — While ML proponents may suggest that agencies can survive a § 706(2)(A) challenge to rulemaking by providing sufficient second-order evidence about how the model was designed,[85] such technical compliance may not always ensure that agency use of ML is in fact transparent, accountable, and rational. First, expecting courts to review the highly technical decisions involved in ML design raises administrability concerns. Courts have long engaged in review of technical agency decisions,[86] but the inscrutability of ML models creates a materially different challenge. Consider, for example, one of the paradigmatic technology cases cited by David Lehr and Professor Cary Coglianese.[87] In *Alaska v. Lubchenco*,[88] the U.S. District Court for the District of Columbia reviewed a traditional model that the National Marine Fisheries Service developed to "extrapolat[e] the negative population trend" of beluga whales in Cook Inlet.[89] To do so, the court discussed various model parameters and assumptions, such as the "constant mortality effect" variable (which accounted for "killer whale predation") as well as other more standard parameters like "growth rates."[90] While the model itself may have relied on complicated statistical techniques, the court had no need to cite formulas or mathematical research. Each individual variable represented a describable, real-world concept, and the model's overall design could be explained. This enabled the court to review the agency's decision on a conceptual level, even if it couldn't understand all the technical details.[91]

---

84 Coglianese & Lehr, *supra* note 24, at 43.

85 *See id.* at 43–44; Averill, *supra* note 41, at 96.

86 *See, e.g.*, Balt. Gas & Elec. Co. v. Nat. Res. Def. Council, Inc., 462 U.S. 87, 89–90 (1983) (evaluating agency assessment of nuclear waste contamination risk); Ethyl Corp. v. EPA, 541 F.2d 1, 10 (D.C. Cir. 1976) (en banc) (evaluating agency assessment of leaded gasoline's risk to public health).

87 Coglianese & Lehr, *supra* note 24, at 45.

88 825 F. Supp. 2d 209 (D.D.C. 2011).

89 *Id.* at 213.

90 *Id.*

91 *See id.* at 223. Similarly, in *Baltimore Gas*, the Court's review of a complex predictive model for nuclear waste contamination ultimately came down to evaluating one variable that could be understood conceptually: an assumption that there would be no risk of contamination by certain types of nuclear waste. 462 U.S. at 90.

Now compare this to the use of an ML model. Although the agency might describe the training data, testing data, and other design decisions, no one — not the agency, the expert witnesses, or the engineers themselves — can describe how real-world variables or concepts relate to the model's outputs.[92] Thus, instead of evaluating the reasonableness of certain real-world assumptions, judges could be faced with inherently mathematical questions like: Is it appropriate for the model to use stochastic gradient descent or root mean squared propagation as its optimization algorithm?[93] Even with expert opinions on both sides, judges may struggle to understand, let alone review these decisions. Compared to judicial review of traditional models, second-order review in the ML context provides less transparency and a weaker check on irrational decisionmaking.

Besides the administrability concerns, an even more fundamental problem remains. Scholars suggest that if an agency proves it did not design a model in an arbitrary or capricious way, the agency action should survive § 706(2)(A) review.[94] However, as will be explored in the remainder of this section, just because a model has been designed in a nonarbitrary and noncapricious manner does not necessarily mean that the model's *output* is nonarbitrary and noncapricious. This matters if the agency then takes action based on that output (as is the case with the FDA and SSA rules). The remainder of the section explores this argument via the case studies.

(a) The FDA Rule. — ML proponents would argue that courts can assess whether the FDA rule is arbitrary and capricious by simply evaluating whether the FDA was arbitrary and capricious in designing the model. Formally, this seems correct. If the agency made a reasonable, rational effort to create a model and confirmed it was running accurately, then it cannot be arbitrary to promulgate a rule generated by such a model, by definition.

However, applying § 706(2)(A) review in this way might do a poor job of filtering out rules that are *practically* "irrational" or "arbitrary." This is because even the most carefully crafted models — models far from being designed arbitrarily — can fail spectacularly when deployed. Consider, for example, Amazon's multiyear attempt to build an ML model that could select top applicants from a set of resumes.[95] After a

---

[92] *See supra* notes 23–29 and accompanying text.

[93] *See, e.g.*, Duk-Sun Shim & Joseph Shim, *A Modified Stochastic Gradient Descent Optimization Algorithm with Random Learning Rate for Machine Learning and Deep Learning*, 21 INT'L J. CONTROL, AUTOMATION, & SYS. 3825, 3825 (2023).

[94] Coglianese & Lehr, *supra* note 24, at 47; *cf.* Averill, *supra* note 41, at 96 ("[A]lgorithms . . . should be able to survive arbitrary and capricious review," since "the agency can still . . . explain its thinking through systemic reasons.").

[95] Jeffrey Dastin, *Insight — Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 10, 2018, 8:50 PM), https://www.reuters.com/article/idUSKCN1MK0AG [https://perma.cc/PR9R-KYLW].

year of development, the team discovered the system was "not rating candidates for software developer jobs and other technical posts in a gender-neutral way."[96] Trained on a disproportionate number of male resumes, the system started "penaliz[ing] resumes" from "graduates of two all-women's colleges" and any resume that included the word "women's."[97] Though the team initially took steps to try to neutralize the gender bias, after four years, they scrapped the project entirely; they simply felt there was no way to "guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory."[98] Beyond the gender bias, the team also struggled with the model making seemingly random recommendations.[99]

An example like this demonstrates the limits of ex ante review in the context of ML. But in fact, those limits are a commonly understood aspect of software engineering more broadly.[100] For example, when a software engineer writes new code, industry standard is to subject the code to a series of complex tests designed to uncover flaws that may have been missed.[101] A second member of the team will also commonly review the code and flag any issues.[102] These are both forms of ex ante review. However, no matter how intensive the initial testing, there will be errors that are uncovered after the fact.[103] To handle these issues, engineers are assigned to "on call" rotations during which they are responsible for triaging any emergency issues that arise and potentially releasing new code to fix the issues.[104] Being "on call" is an especially stressful aspect of being a software engineer, and it is uncommon for no

---

[96] *Id.*

[97] *Id.*

[98] *Id.*

[99] *Id.* Cameron Averill acknowledges that ML models "can change dynamically over time," but suggests that as long as the agency "explain[s] why it does not expect changes to be dramatic," there should be no problem with arbitrary and capricious review. Averill, *supra* note 41, at 97. However, just because engineers reasonably *expect* no drastic changes doesn't mean they won't occur.

[100] Seokjoo Andrew Chang et al., *Debugging Debugging*, J. MULTIDISCIPLINARY RSCH., Spring 2019, at 51, 52.

[101] *See* Sten Pittet, *The Different Types of Software Testing*, ATLASSIAN, https://www.atlassian.com/continuous-delivery/software-testing/types-of-software-testing [https://perma.cc/E2AL-6BUC].

[102] This process is typically described as making a "pull request." *About Pull Requests*, GITHUB DOCS, https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/proposing-changes-to-your-work-with-pull-requests/about-pull-requests [https://perma.cc/AJ9L-57EL].

[103] Even in traditional software engineering, it's essentially impossible to catch all bugs at the outset. *See Is It Possible to Reach Absolute Zero Bug State for Large Scale Software?*, STACKEXCHANGE: SOFTWARE ENGINEERING (July 20, 2019, 12:21 PM), https://softwareengineering.stackexchange.com/questions/195571/is-it-possible-to-reach-absolute-zero-bug-state-for-large-scale-software [https://perma.cc/4HWC-2JLV].

[104] *See* Stefan Nothaas, *Being On-Call as a Software Engineer — A Challenging and Fast Learning Experience*, TRIVAGO (Jan. 12, 2022), https://tech.trivago.com/post/2022-01-12-engineeroncall [https://perma.cc/5AJY-NDSS].

issues to arise.[105] In other words, even in traditional software engineering, the ability to catch and resolve issues ex post is important to ensuring high-functioning algorithms. These challenges are only heightened in the context of ML models, due to their enhanced complexity and unintuitive nature.[106]

In sum, guaranteeing that the FDA did not act arbitrarily or capriciously in designing and testing its ML model ex ante is not enough to ensure that the resulting rules generated by the model will be nonarbitrary. Because it is genuinely difficult to anticipate all issues that might arise from an ML model ex ante, one could make completely reasonable second-order decisions about the model, but still end up with highly flawed outputs. Even more concerning, while some models will fail spectacularly, others may fail quietly.[107] For example, it might be obvious when an Amazon recruitment model recommends only male candidates; it would be far less obvious if the FDA model began making incorrect predictions about the health risks of unapproved drugs.

Finally, second-order review of ML rulemaking might also fail to effectively police the relevant factors requirement.[108] Suppose, for example, that the FDA's organic statute requires that three specific factors be considered when regulating drugs: F_1, F_2, and F_3. On a second-order level, the FDA might formally meet this requirement by ensuring that the model is trained on, and takes as input, data related to each of the three factors.[109] However, when the FDA actually runs the model on a new drug, there is no way to know *which* input data is important to the model's decision. The model might actually be placing 99.99% of its weight on only two of the factors[110] — or it may be considering a completely different, prohibited factor, F_4, that it was able to infer from the training data.[111] As before, second-order constraints cannot ensure the model is acting nonarbitrarily on a first-order level.

(b) The SSA Rule. — In defending its decision to use ML to prioritize cases, the SSA may provide many different types of information: validation tests showing the accuracy and efficiency of the model, the

---

[105] *See* Jiaqi Liu, *How to Build a Successful On-Call Culture: A Guide for Engineering Teams*, BUTTON, https://www.usebutton.com/post/fostering-a-strong-engineering-on-call-culture [https://perma.cc/3GTV-N4XE].

[106] *See supra* notes 30–32 and accompanying text.

[107] Ayush Patel, *The Subtle Art of Fixing Silently Failing ML Models*, CENSIUS, https://censius.ai/data-science-festival-how-to-fix-ml-models [https://perma.cc/5KKK-NUFQ] (describing "[s]ilent failure" as "when model performance gradually degrades . . . without showing any apparent signs of failure" for a period of time).

[108] *See supra* notes 68–69 and accompanying text.

[109] *See* Averill, *supra* note 41, at 96 (making this argument).

[110] Sarid & Ben-Zvi, *supra* note 2, at 384 (describing how information about an ML's design does not provide insight on the reasons relied on in the "*specific decision* under examination").

[111] The Amazon recruiter model is a more concrete example of this. While the engineers did not want the model to rely on discriminatory factors (and explicitly prevented the model from relying on certain terms), they could not ensure that the model would not find other ways to effectively rely on such factors. *See* Dastin, *supra* note 95.

policy benefits of using the model, the training data and design of the model, and so on. In this way, the agency should be able to prove that it considered all relevant factors and made a rational, informed decision in choosing to introduce ML to the claims process. However, as with the FDA rule, it is unclear that this type of arbitrary and capricious review would be sufficient to prevent the model from *acting* in an arbitrary manner once deployed.

ML proponents would argue that reviewing second-order information about the model should be a sufficient safeguard. First, if the model is going to act in an arbitrary way once released, there must be some defect that can be uncovered when the model is subject to notice and comment; this Note has already discussed why this is not necessarily the case.[112] Second, ML proponents could argue that while second-order review cannot *ensure* that the model won't go rogue, such assurance also isn't typically required for rules relying on human decisionmakers. Suppose, for example, that instead of adopting an ML model, the SSA issued a parallel human-version of its ML rule: All cases would go through a high-performing administrator, who would predict the likelihood of claim approval based on consideration of various factors (severity of disability, economic need, and so forth). They would then approve or "slow track" outlier cases. Like the ML rule, ensuring that the factors and surrounding procedures are not arbitrary or capricious cannot guarantee that the human decisionmaker will not act arbitrarily. But rules granting broad discretion are standard fare in the administrative state and generally aren't deemed arbitrary and capricious.[113]

However, there are at least two reasons to believe that human decisionmakers are less likely to act arbitrarily than an ML model, even in the context of highly discretionary rules. First, humans operate in the shadow of review. While rules often authorize human discretion, they also often check that discretion by requiring the decisionmaker to provide some sort of explanation or by providing for an appeals process.[114] And even when none of those requirements are in place, humans still have managers who might review their work. Because humans know that they may be asked to explain themselves, they are more likely to operate in a way that could be justified and considered reasonable.[115]

---

[112] *See supra* notes 79–111 and accompanying text.

[113] *See, e.g.*, 12 C.F.R. § 5.20(f) (2024) (listing multiple open-ended factors the Office of the Comptroller of the Currency should consider when granting bank charters); 32 C.F.R. § 147.2 (2024) (same for the Office of the Secretary of Defense in making security clearance determinations).

[114] *See* Strandburg, *supra* note 23, at 1864 ("Reason giving is a core requirement in conventional decision systems precisely *because* human decisionmakers are inscrutable and prone to bias and error . . . .").

[115] *See id.* at 1868, 1868–69; *see also* Joseph Burgo, *Why Shame Is Good*, VOX (Apr. 18, 2019, 1:30 PM), https://www.vox.com/first-person/2019/4/18/18308346/shame-toxic-productive

In contrast, the ML model is constrained in no such way. It feels no shame and, by definition, cannot describe the reasoning behind any of its decisions. For these reasons, human SSA decisionmakers may be practically more constrained than the ML model.

Second, human decisionmakers may be constrained by broader concepts like "fairness" or the SSA's general mission.[116] The SSA's ML model has no independent notion of what is wrong with only slow tracking cases with the name "Daniel" — just as the Amazon recruiter model had no sense of what was wrong with only recommending male candidates. In contrast, a good faith human decisionmaker would never get to the point of "slow tracking" only claimants with the name "Daniel." Their broader understanding about the world — the purpose of disability claims, human naming conventions, fairness — serves as a sort of backstop, signaling that there must be something wrong with their decisionmaking process if they are only flagging claims with one name. ML proponents might counter that while humans may not be irrational in the specific way that machines are, they can be forgetful, biased, and irrational in their own ways.[117] However, because the SSA's ML model is ultimately trained on existing human-made decisions,[118] it is not obvious that such human biases would be eliminated from the ML model either.[119] Indeed, in the worst case scenario, an ML model might freeze these biases and preserve them indefinitely.[120]

2. "Extra" Deference for Technical Cases. — In arguing that agency use of ML can survive arbitrary and capricious review, scholars highlight that courts are especially deferential "when the reasons that an agency offers" for its action "depend on highly technical matters," citing cases like *Baltimore Gas* as precedent.[121] Though ML rulemaking may seem like a paradigmatic example of a "highly technical matter," the underlying justifications for *Baltimore Gas* may not always extend to

---

[https://perma.cc/C6Q5-4M4L] ("[O]ur personal sense of shame may also help us meet our own expectations and live up to our values."); Sarid & Ben-Zvi, *supra* note 2, at 391 ("Agencies thus make decisions while knowing that they must be transparent about their reasoning . . . .").

116 *Mission and Structure*, SOC. SEC. ADMIN., https://www.ssa.gov/agency [https://perma.cc/L83V-JG93].

117 *See* Cary Coglianese & Alicia Lai, *Algorithm vs. Algorithm*, 71 DUKE L.J. 1281, 1281 (2022) ("[H]uman decision-making suffers from memory limitations, fatigue, cognitive biases, and racial prejudices, among other problems."). This idea is also a fundamental principle behind behavioral economics. *See* Saurabh Jha & Adam Powell, *A (Gentle) Introduction to Behavioral Economics*, 203 AM. J. ROENTGENOLOGY 111, 111 (2014).

118 *See supra* notes 51–52 and accompanying text.

119 *See* James Manyika et al., *What Do We Do About the Biases in AI?*, HARV. BUS. REV. (Oct. 25, 2019), https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai [https://perma.cc/798W-QEQM] (describing how biases can specifically become "bak[ed] in" to AI models).

120 *See* IBM Data and AI Team, *Shedding Light on AI Bias With Real World Examples*, IBM (Oct. 16, 2023), https://www.ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples [https://perma.cc/K7GT-SA3H].

121 Coglianese & Lehr, *supra* note 24, at 28.

ML rulemaking. This section explores such caveats through the lens of the FDA and SSA rules.

In traditional technical cases, one reason to give special deference to agencies is that they are operating within their "area of special expertise."[122] But this may not be true in certain forms of ML rulemaking. Imagine the FDA rule had been generated, not by an ML model, but by human administrators. In creating the rule, the administrators utilized their experience from evaluating dozens of drug trials and their understanding of the adverse events reporting system. A generalist judge asked to review the FDA's defense of the rule would have no comparable experience with drug trials and, consequently, should defer. However, when generating the same rule via ML, the FDA does *not* seem to operate within its area of expertise. If the ML proponent's argument[123] is taken on its own terms, the appropriate focus for arbitrary and capricious review would be second-order questions about the FDA's ML model: whether it was well designed, trained, and tested.[124] But these questions implicate pure computer science, not just the regulation of drugs. The FDA has no subject matter expertise in or intuition for these kinds of questions.

There are at least two counterarguments that might be made. First, one might contest that the agency lacks expertise. Even if the FDA has no institutional experience with ML, whoever built the model had expertise and thus deserves deference. Certainly, some amount of expertise is required even to create an ML model. But simply having experience using a particular technical tool seems quite different from the deep subject-matter expertise that merited deference in cases like *Baltimore Gas*.[125] For example, any general statistician may be able to create a model based on the FDA's databases, but this skill wouldn't seem to put them on par with FDA experts who have spent decades analyzing pretrial drug tests and predicting potential risks. Similarly so for an ML "expert" who has experience building tech advertising models but is asked to build a model on the FDA's adverse event data.

Second, one might argue that even if ML design is outside an agency's core expertise, the agency still has more expertise than a generalist judge. If the "extra" deference of *Baltimore Gas* is justified by fear of judicial overreach in areas where courts have little understanding, then it should be equally applicable in the context of an ML-generated rule like the FDA's. To the extent that arbitrary and capricious review is conceived of as a judge searching for substantive

---

122 Balt. Gas & Elec. Co. v. Nat. Res. Def. Council, Inc., 462 U.S. 87, 103 (1983).

123 *See supra* section II.B, pp. 1830–31.

124 Coglianese & Lehr, *supra* note 24, at 43–44.

125 *See Balt. Gas*, 462 U.S. at 98–100 (describing the complex, detailed analyses conducted by the agency in estimating the risk of nuclear leakage).

errors of judgment, as Judge Leventhal argued,[126] this argument seems correct. Between the court and the agency, the court is unlikely to successfully pinpoint substantive errors that the FDA made in developing its model. However, to the extent that § 706(2)(A) review is viewed as a court's responsibility to ensure that the agency has really grappled with and responded to arguments made by the public,[127] the FDA's relative lack of expertise in ML would still seem to matter. While a generalist judge may not have more expertise than the FDA with respect to designing ML models, a computer science professor submitting a comment might. If anything, the difference in expertise between the public and the FDA makes it even *more* important that courts hold the agency accountable for responding to all material comments, since these comments are more likely than usual to have caught significant flaws (as compared to rules in which the FDA utilizes its core expertise).

A separate justification for giving agencies special deference when making highly technical decisions is that they often do so under genuine uncertainty. As Professor Adrian Vermeule describes, "[a]gencies frequently encounter novel problems at the frontiers of scientific and technical knowledge, such that even expert probability assessments are unreliable and real uncertainty is pervasive."[128] Under these conditions, agencies *must* make a decision — "some choice or other is inescapable, legally mandatory, or both" — but "no first-order reason can be given."[129] For example, in *Baltimore Gas*, the Nuclear Regulatory Commission engaged in a complex predictive risk analysis to determine the "environmental effects of a nuclear powerplant's fuel cycle."[130] As both the Commission and Court acknowledged, the assumption under review was "surrounded with uncertainty"[131] and "rigorous verification of long-term risks for waste repositories [was] not possible."[132]

Coglianese and Lehr imply that this justification naturally extends to ML rulemaking, stating that extra deference is given "when the *reasons* that an agency offers" for its action "depend on highly technical matters."[133] But to be more precise, the focus of *Baltimore Gas* does not seem to be the complexity of the *tools* used, but the complexity of the *question* posed to the agency: whether nuclear waste might leak into the environment over the course of many years.[134] The difficulty of this

---

126 MANNING & STEPHENSON, *supra* note 5, at 988–91 (citing Greater Bos. Television Corp. v. FCC, 444 F.2d 841, 851–52 (D.C. Cir. 1970)).

127 *See* United States v. N.S. Food Prods. Corp., 568 F.2d 240, 252 (2d Cir. 1977) ("It is not in keeping with the *rational* process to leave vital questions, raised by comments . . . of cogent materiality, completely unanswered." (emphasis added)).

128 VERMEULE, *supra* note 70, at 153.

129 *Id.* at 129.

130 *Balt. Gas*, 462 U.S. at 91.

131 *Id.* at 96.

132 *Id.* at 104.

133 Coglianese & Lehr, *supra* note 24, at 28 (emphasis added).

134 *See Balt. Gas*, 462 U.S. at 104–05.

question is what made it impossible for the agency to give first-order reasons for its prediction, which in turn justified extra deference from the court.[135]

Practically, complex tools often go along with complex questions. However, the versatility of ML models means that while they are sometimes used to answer complex, predictive questions,[136] other times they are not. This is especially so when ML is used to automate repetitive but relatively simple tasks.[137] For example, imagine that the SSA decided to use an ML model not to predict claim outcomes, but to extract specific data from a claimant's documents, such as the date of the claimant's last doctor's appointment. Unlike in *Baltimore Gas*, there is no uncertainty inherent to these questions; there is a right answer that can be determined with certainty if a human simply reviews the materials.

In these circumstances, the "uncertainty" rationale underlying *Baltimore Gas* does not extend to ML rulemaking. But to be clear, this does not mean that the agency action must be found arbitrary and capricious. In the hypothetical example above, the SSA might have had very important reasons for adopting the model; perhaps it would significantly reduce reversals caused by erroneous human review of documents. In such cases, courts should require the agency to provide valid reasons justifying its action, just as they would for any conventional rule. In other words, the use of a technical tool like ML does not guarantee that the underlying problem involves genuine uncertainty, and thus, should not automatically guarantee heightened deference.[138]

* * *

This Part has explored how arbitrary and capricious review might apply to different forms of ML rulemaking. Reviewing courts should keep two key points in mind. First, while agencies may be able to formally prove that ML rulemaking is nonarbitrary by providing sufficient second-order data — about how the model was designed, tested, and why it makes good sense to use it — this cannot ensure the model's outputs will be rational. The complexity of designing ML models and their ability to degrade over time means they may end up acting erratically, even if designers made the most reasonable decisions possible while

---

135 VERMEULE, *supra* note 70, at 133; *see also id.* at 182–83.

136 For example, when the FDA seeks to use an ML model to predict negative health effects on children, it clearly *does* face a complex, predictive question.

137 *See, e.g.*, Kevin Casey, *How to Explain Robotic Process Automation (RPA) in Plain English*, ENTERPRISERS PROJECT (July 30, 2020), https://enterprisersproject.com/article/2019/5/rpa-robotic-process-automation-how-explain [https://perma.cc/4LQ3-PP9X] (describing how ML can be used to "automat[e] some of the most mundane and repetitive computer-based tasks").

138 For an exploration of how to determine the appropriate level of deference, see Cade Mallett, *Judicial Deference to Agency Action Based on AI*, 32 CATH. U. J.L. & TECH. 37, 54–74 (2023) (suggesting factors courts should consider when determining degree of deference to give to AI-based agency actions).

building them. Moreover, this risk is greater in ML rulemaking than in conventional rules, even those involving significant discretion.

Second, the historic deference given to especially technical agency decisions should not be blindly extended to all ML rulemaking. The two common justifications for such deference — that the agency was operating (1) within its special area of expertise and (2) under genuine uncertainty — will not always hold true in ML rulemaking.

In sum, the tensions raised in this section demonstrate that § 706(2)(A) will sometimes fail to achieve its underlying goal in the context of ML rulemaking: to promote transparent, accountable, and rational decisionmaking.

III. POTENTIAL SOLUTIONS

If the existing requirements of § 706(2)(A) may fail to guarantee a robust rulemaking process when agencies use ML, what requirements would? While a comprehensive answer is beyond the scope of this Note, the legal analyses above suggest at least one possibility: a greater focus on agency obligations *after* an ML rule is finalized.

Most APA requirements for rulemaking focus on the time before a rule is adopted. Notice-and-comment requirements establish what steps must be taken before the rule goes into effect;[139] similarly, § 706(2)(A) review focuses on the agency's reasoning at the time the rule was finalized.[140] This is sensible for conventional rules that can be more clearly defined in advance and that do not typically evolve over time. However, establishing agency obligations that apply after a rule is adopted may be a better fit for ML rulemaking, given the inscrutable and evolving nature of ML models.

Consider two potential ex post obligations for ML rulemaking: an extended notice obligation and an extended arbitrary and capricious challenge. An extended notice obligation[141] would require agencies to provide ongoing transparency into ML models used in rulemaking, even after a final rule is promulgated. For example, agencies might be required to provide annual reports on the current models in use, their error rates, descriptive statistics on the models' decisions, and any anomalous outputs that were discovered. Working in tandem, an extended arbitrary and capricious challenge would allow the public to challenge the agency's decision to *continue* using an ML model (rather than simply its initial decision to promulgate the ML rule). In evaluating such a claim, courts would consider not only the record at the time the rule was

---

139 *See* 5 U.S.C. § 553(b)-(c).

140 *See* SEC v. Chenery Corp. (*Chenery I*), 318 U.S. 80, 87 (1943); MANNING & STEPHENSON, *supra* note 5, at 1049 (describing how *Chenery I* applies in the context of hard look review).

141 *See supra* notes 54–55 and accompanying text (summarizing the APA's existing notice requirements).

finalized, but also the ongoing "status reports" produced by the agency to date.

Imposing these ex post obligations would get to the core of the issues raised in section II.C. As discussed, a fundamental problem with existing § 706(2)(A) review of ML rulemaking is that even when a model is not designed in an arbitrary or capricious manner, over time, it may begin functioning in an arbitrary manner. Recall the hypothetical degradation of the SSA model, which, though reasonably designed, begins slow tracking only claimants named "Daniel." Annual reports on the SSA's model would reveal the strange pattern of "slow track" decisions. And if the SSA decided to continue using the rule, a claimant could challenge it as arbitrary and capricious. By shifting the focus of arbitrary and capricious review from an ex ante to an ex post perspective, the proposed requirements would take advantage of the fact that it is far easier to catch absurd results generated by ML models ex post than to predict them.

Even if these obligations are useful in theory, how might they be implemented? The most obvious path would be for Congress to amend the APA or pass new legislation imposing greater ex post obligations on agency use of ML. This legislation could explicitly establish reporting obligations, define criteria for when such obligations are triggered, and extend arbitrary and capricious review. While perhaps the most direct option, it also is the least practical, given the gridlocked state of Congress.[142]

A second option would be to leverage existing APA provisions, such as the notice-and-comment process. Suppose, for example, that after the SSA gave notice of its ML rule, a comment was submitted highlighting the danger of ML model degradation and requesting the agency to commit to providing ongoing status reports while the model is in use. The agency might always choose to reject this proposal, but it would at least need to provide a reasonable defense; moreover, its decision to decline such obligations could itself be challenged as arbitrary and capricious. Similarly, rulemaking petitions might allow the public to challenge an agency's decision to continue using an ML model as arbitrary and capricious.[143] Under § 553(e), "interested person[s] [have] the right to petition for the . . . repeal of a rule."[144] Denials of such petitions should generally be accompanied by "a brief statement of the grounds for

---

[142] *See* Eric McDaniel, *Congress Passed So Few Laws This Year that We Explained Them All in 1,000 Words*, NPR (Dec. 22, 2023, 5:00 AM), https://www.npr.org/2023/12/22/1220111009/congress-passed-so-few-laws-this-year-that-we-explained-them-all-in-1-000-words [https://perma.cc/U2ZF-SSK5].

[143] *Cf.* Engstrom & Ho, *supra* note 15, at 853 (discussing how arbitrary and capricious review might be used to impose benchmarking requirements on agencies).

[144] 5 U.S.C. § 553(e).

denial"[145] that "explain[s] why [the agency] decided to act as it did."[146] These denials are subject to judicial review.[147] In fact, judicial review of petition denials has been explicitly identified as the correct way to challenge a rule that has become "obsolete over time."[148]

The main strength of this approach is that it is immediately implementable, but it also has key limitations. First, while agencies must respond to all material public comments,[149] this would still seem to leave substantial room for the agency to avoid ongoing notice requirements, by arguing, for example, that these commitments would be overly burdensome or invasive.[150] Even if the agency agreed to provide some ongoing transparency, it might commit to providing so little data that the public could not mount meaningful challenges to the rule. Successfully policing the line here would seem to be difficult for courts. Second, even though agency denials of rulemaking petitions are subject to arbitrary and capricious review, courts have also emphasized that the standard of review should be "'extremely limited' and 'highly deferential.'"[151] In extreme cases of degradation, courts might be willing to require the agency to stop using the ML model; but in these cases, the agency itself would likely agree. In more contested cases, this extremely deferential standard of review might significantly weaken the public's ability to challenge the agency's continued use of ML.

CONCLUSION

Whether one is hopeful or pessimistic about the future of ML, federal agencies' reliance on such models only seems likely to grow. It is critical, then, that our legal frameworks effectively guide agency use of ML. This Note suggests that in some cases, ML rulemaking will raise no novel issues with APA compliance. In others, formalistic compliance may be possible but will fail to achieve the APA's goals of rationality and transparency as successfully as in conventional rulemaking. Expanded ex post requirements may be one way to fill this gap. But regardless of the specific solution, what *is* clear is the need to grapple with the genuinely revolutionary aspects of ML and to do so quickly.

---

145 *Id.* § 555(e).

146 Butte County v. Hogen, 613 F.3d 190, 194 (D.C. Cir. 2010) (citing Tourus Recs., Inc. v. Drug Enf't Admin., 259 F.3d 731, 737 (D.C. Cir. 2001)).

147 *See* Massachusetts v. EPA, 549 U.S. 497, 527–28 (2007) (quoting Nat'l Customs Brokers & Forwarders Ass'n of Am., Inc. v. United States, 883 F.2d 93, 96 (D.C. Cir. 1989)); *see also* HICKMAN & PIERCE, *supra* note 64, § 4.10.

148 HICKMAN & PIERCE, *supra* note 64, § 4.10; *see* Auer v. Robbins, 519 U.S. 452, 459 (1997).

149 *See* HICKMAN & PIERCE, *supra* note 64, § 5.4; *see also* United States v. N.S. Food Prods. Corp., 568 F.2d 240, 252 (2d Cir. 1977).

150 For example, agencies might argue that the data may not be disclosed given privacy, trade secret, or other legal concerns. *See* Averill, *supra* note 41, at 97. *But see* Coglianese & Lehr, *supra* note 24, at 48–49 (questioning the strength of these arguments).

151 Massachusetts v. EPA, 549 U.S. at 527–28 (quoting *Nat'l Customs Brokers & Forwarders*, 883 F.2d at 96).