Catholic University Journal of Law and Technology

Volume 32
Issue 1 *Fall 2023*

Article 4

2023

Judicial Deference to Agency Action Based on AI

Cade Mallett
New York University School of Law

JUDICIAL DEFERENCE TO AGENCY ACTION BASED ON AI

Cade Mallett *

When reviewing agency action for arbitrariness, courts must initially determine how "hard" of a look to take at the substance of agency action. The increasing use of AI as a basis for agency action threatens to complicate this threshold analysis significantly, as agencies and courts are both commonly lacking in significant expertise creating and reviewing AI. While it is common

* J.D., NYU Law 2024 (expected). I would like to express my deepest gratitude to Professor Catherine Sharkey of NYU, whose guidance and feedback throughout the process of writing and editing this article were invaluable. Thanks also to the Privacy Research Group at NYU and its faculty director, Professor Katherine Strandburg, to whom I presented the article and who gave many thoughtful notes and suggestions. Thank you to Soumya Kandukuri for thoughtful comments and feedback. Finally, thank you to the staff of the *Catholic University of America Journal of Law and Technology* for their thorough edits and revisions.

for lower courts to rotely determine they are entitled to "hard look" review of agency action, the Supreme Court's precedent in this area is decidedly more deferential, requiring a case-by-case assessment of the extent to which an agency leverages its substantive expertise in taking the action. Leveraging both the Court's expertise-based analysis and a review of the policy considerations underlying the decision to grant deference, this paper contributes a framework for courts to use in choosing the level of deference to grant agency action based on AI.

The paper proceeds as follows. Part I describes the current landscape of agency use of AI and the hazards that await if agencies are not careful in their adoption of this new technology. Part II considers the existing law surrounding decisions to defer to agency action, rejecting a common tendency to default to hard look review in favor of an approach tailoring deference to the substantive expertise the agency displays per the Court's holding in *Baltimore Gas*. Part III identifies factors that demonstrate the substantive expertise an agency brings to bear through use of AI and analyzes their impact on the deference decision based on the deference inquiry described in Part II, as well as policy concerns underlying the determination.

I. AGENCIES AND AI

A. What is AI?

Computer science scholars define artificial intelligence as "the study of agents that receive percepts from the environment and perform actions."[1] This definition is intentionally broad to avoid imposing artificial restrictions on the ways that we can design systems capable of thought.[2] To put the issue more practically, a 2020 report on AI drafted for the Administrative Conference of the United States (the "ACUS report") limited its scope to "the most recent forms of machine learning, which train models to learn from data."[3] I will adopt the same scope, which, according to the report, most notably encapsulates

---

1 STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH (4th ed. 2021).

2 For an example of some useful subdivisions, Russel and Norvig create distinct categories of systems based on humans and those based on ideal rational reasoning. *Id.* at 1–6. They further subdivide those categories based on whether the system primarily thinks either as a human or ideal reasoner, or primarily acts in that manner. *Id.*

3 DAVID FREEMAN ENGSTROM ET AL., GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 12 (2020) (report to the Admin. Conf. of the U.S.), https://www.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf [https://perma.cc/5S6Q-XC3S].

"'supervised learning,' where 'training data' is used to develop a model with features to predict known 'labels' or outcomes, and 'unsupervised learning,' where a model is trained to identify patterns in data without labels of interest."[4] I will exclude many popular AI technologies, like rule-based or "expert" systems, and more conventional forms of statistics.

Even within this narrowed domain, AI has taken many different forms and uses. In addition to the supervised/unsupervised division, AI can accept and produce various forms of input and output (images, text, numbers, or custom-defined categories or other types); can adapt differently ("online" AI can adapt to new data on the fly, while "offline" AI are sedentary and do not change their output after training is finished); and can be either more or less explainable (while some techniques are designed to produce outcomes humans can understand and reproduce, others develop their predictive power directly from the ability to sleuth out patterns beyond human comprehension).[5] In addition, AI tools assist their users in different ways, ranging from producing a mere suggestion to producing an authoritative decision, and from providing a small part of an answer to providing a complete solution.[6] Prevalent AI uses include computer vision, natural language processing ("NLP"), sentiment analysis, and text to category conversion.[7]

B. Agency Use of AI

Like AI, the federal government is expansive. The scope of "federal agencies" within this paper includes executive departments, their sub-components, and independent agencies. The ACUS report performed a canvas of 142 federal administrative departments and agencies to assess trends in agency use of AI.[8] Agencies are using algorithms for internal operational improvements, civil and law enforcement, formal and informal adjudication, risk monitoring and analysis, data monitoring, and communication with the public.[9] While 45 percent of canvassed agencies planned, piloted, or implemented AI techniques, use cases were heavily concentrated in just a few agencies, with 7 percent of the agencies producing 70 percent of the AI tools in the report.[10] The report

---

4 *Id.*

5 *Id.* at 10–11, 19.

6 *See* CARY COGLIANESE, A FRAMEWORK FOR GOVERNMENTAL USE OF MACHINE LEARNING 3–4 (Dec. 8, 2020) (report to the Admin. Conf. of the U.S.), https://www.acus.gov/sites/default/files/documents/Coglianese%20ACUS%20Final%20Report.pdf.

7 ENGSTROM ET AL., *supra* note 4, at 27, 40, 42, 61–62.

8 *Id.* at 6.

9 *Id.* at 13.

10 *Id.* at 16.

classified only 12 percent of tools deployed by agencies as high-sophistication, possibly underscoring the technology gap between the public and private sectors.[11] In addition, 53 percent of agency AI tools were made in-house.[12]

The ACUS report also conducted case studies on particular uses of AI in federal agencies. The Securities and Exchange Commission ("SEC") uses an array of AI tools in regulatory enforcement, including to identify violations of federal securities laws like fraud in accounting and financial reporting and trading-based market misconduct.[13] Customs and Border Protection ("CBP") has integrated facial recognition technology and passenger risk scoring tools into its law enforcement strategies, though the agency has relied primarily on contractors for these tools.[14] The Social Security Administration ("SSA") has adopted AI tools into at least three stages of formal adjudications for disability claims: automatically clustering similar cases for assignment to a single administrative law judge ("ALJ") to aid in specialization; predicting the likelihood of success of appeals to fast-track the most promising cases; and performing NLP on written decisions to determine whether draft decisions deviate from agency policy or are internally inconsistent.[15] The Patent and Trademark Office ("PTO") has adopted a third-party machine learning ("ML") model for classifying patent applications into subject areas, as well as in-house AI tools to search for prior art in patent prosecutions, classify trademark

---

11 *Id.* at 20. An alternate reading of this statistic is that the federal government is constrained to use low-sophistication AI regardless of its technical capacity due to other requirements not felt in the private sector, like the need for explainability. Since we do not know the baseline for comparison, the percent of private sector AIs which are sophisticated, it may also be that 12% of AIs being sophisticated is standard. The tendency both in private and public sectors is likely to use the simplest tool capable of the use case, so a low sophistication level may be reasonable.

12 *Id.* at 7.

13 *Id.* at 23–24; *DERA - Office of Risk Assessment*, SEC. & EXCH. COMM'N, https://www.sec.gov/page/dera_ora_page (Dec. 1, 2021); Mary Jo White, *Remarks at the International Institute for Securities Market Growth and Development*, SEC. & EXCH. COMM'N (Apr. 8, 2016), https://www.sec.gov/news/statement/statement-mjw-040816.html.

14 ENGSTROM ET AL., *supra* note 4, at 30–31; Marcy Mason, *Biometric Breakthrough: How CBP Is Meeting Its Mandate and Keeping America Safe*, CBP: FRONTLINE, https://www.cbp.gov/sites/default/files/assets/documents/2018-Jun/Frontline%20-%20Vol9%20Iss3%20-%20Web.pdf (last visited Nov. 24, 2023); DEP'T. HOMELAND SEC., DELIVERY ORDER HSBP1017J00203, USA SPENDING, https://www.usaspending.gov/award/CONT_AWD_HSBP1017J00203_7014_HSHQDC13D00020_7001; *Risk Prediction Program*, U.S. DEP'T HOMELAND SEC., SCI. & TECH. (2014), https://www.dhs.gov/sites/default/files/publications/Risk%20Prediction-508_0.pdf.

15 ENGSTROM ET AL., *supra* note 4, at 39–40; FELIX F. BAJANDAS & GERALD K. RAY, ADMIN. CONF. U.S., ELECTRONIC CASE MANAGEMENT SYSTEMS IN FEDERAL AGENCY ADJUDICATION 2 (2018), https://www.acus.gov/sites/default/files/documents/2018.06.25%20eCMS%20post-Plenary%20FINAL_0.pdf; SOC. SEC. ADMIN., UPDATED COMPASSIONATE AND RESPONSIVE SERVICE (CARES) AND ANOMALY PLAN 12–13 (2017).

applications, and search for prior trademarks.[16] The Food and Drug Administration ("FDA") has piloted the use of AI/ML tools to perform regulatory analysis by building a database of text-based reports of adverse drug events as well as NLP tools to analyze those reports, among other things.[17] The Federal Communications Commission ("FCC") and Consumer Financial Protection Bureau ("CFPB") have implemented AI tools for public engagement.[18] The FCC hired a consulting firm to perform sentiment analysis on comments submitted online as part of notice and comment for a proposed regulation on net neutrality, while the CFPB uses NLP to process, prioritize, and respond to consumer complaints.[19] Finally, the United States Postal Service ("USPS") has developed prototype autonomous vehicles to aid postal service workers in delivery by permitting them to sort mail while the vehicle is in motion.[20]

C. Challenges for Agencies in Adopting AI

To ground the discussion of review of agency action based on AI, we must consider the challenges facing agencies adopting AI, and the risks they pose. The ACUS report highlights six challenges.[21] First, agencies will struggle to develop the technical capacity to make and use AI, facing challenges, including the difficulty of establishing a data pipeline, insufficient resources to attract talent, and regulatory barriers making AI development and adoption more

---

16 ENGSTROM ET AL., *supra* note 4, at 48–49; U.S. PAT. & TRADEMARK OFF., PTOC-016-00, PRIVACY IMPACT ASSESSMENT FOR THE SERCO PATENT PROCESSING SYSTEM (PPS) 1 (2018); Arthi Krishna et al., *Examiner Assisted Automated Patents Search*, AAAI FALL SYMP. SERIES: COGNITIVE ASSISTANCE IN GOV'T & PUB. SECTOR APPLICATIONS 153 (2016); Coding of Design Marks in Registrations, 75 Fed. Reg. 81, 587–88 (Dec. 28, 2010), https://www.regulations.gov/document?D=PTO-T-2010-0090-0001; *Emerging Technologies in USPTO Business Solutions*, U.S. Patent & Trademark Off. 14, 18 https://www.wipo.int/edocs/mdocs/globalinfra/en/wipo_ip_itai_ge_18/wipo_ip_itai_ge_18_p5.pdf (last visited Nov. 24, 2023).

17 ENGSTROM ET AL., *supra* note 4, at 55; *see Postmarketing Surveillance Programs*, U.S. FOOD & DRUG ADMIN. (Apr. 2, 2020), https://www.fda.gov/drugs/surveillance/postmarketing-surveillance-programs; QAIS HATIM ET AL., U.S. FOOD & DRUG ADMIN., MODELING AND TEXT ANALYSIS TO EMPOWER FAERS ADVERSE EVENT ASSESSMENT 5–6 (2018).

18 ENGSTROM ET AL, *supra* note 4, at 60–61.

19 ENGSTROM ET AL, *supra* note 4, at 60–61; EMPRATA, COMMENTS ANALYSIS FCC RESTORING INTERNET FREEDOM DOCKET 17-108, 2 (2017); STRATEGIC PLAN, BUDGET, AND PERFORMANCE PLAN AND REPORT, CONSUMER FIN. PROT. BUREAU 47 (Mar. 2014), https://files.consumerfinance.gov/f/strategic-plan-budget-and-performance-plan-and-report-FY2013-15.pdf.

20 ENGSTROM ET AL., *supra* note 4, at 66–67; U.S. POSTAL SERVICE, OFF. OF THE INSPECTOR GEN., AUTONOMOUS VEHICLES FOR THE POSTAL SERVICE 7 (2017), https://www.uspsoig.gov/sites/default/files/reports/2023-01/RARC-WP-18-001.pdf.

21 ENGSTROM ET AL., *supra* note 4, at 6–7.

difficult.[22] Agencies will need to build internal capacity to realize the promises and benefits of algorithmic governance.[23] Procurement is an incomplete solution to the lack of internal capacity because contracting for custom goods like agency-specific AIs can invite corner-cutting by the manufacturer.[24] Agencies will need to invest in their technology infrastructure and in-house human capital to produce technically effective and legally compliant AI tools, which may have the added benefit of promoting public accountability and the transparency of the agency's AI strategies as well.[25]

Second, agencies will struggle to promote transparency and accountability due to the inherent opacity of many advanced AI techniques.[26] The use cases explored in the ACUS report show hard trade-offs between accountability and efficacy.[27] For example, the SEC's use of algorithmic enforcement tools has prompted calls to deliberately impair the tools' predictive accuracy to increase its explainability.[28] Agencies will also face challenges creating tools that conform to the existing structure of administrative law. The report suggests updating the Administrative Procedure Act ("APA"), judicially or legislatively, to foster more transparency through ex ante review of algorithmic tools by leveraging standard procedural review processes like notice and comment.[29]

Third, agencies will struggle to weed out bias, disparate treatment, and disparate impact from AI, especially where they have limited insight into the algorithms' functioning.[30] Protected characteristics, even when excluded from an algorithm's training data, can be inferred with high accuracy.[31] Moreover, even if protected characteristics could be completely excluded from an AI's decision making, it is often mathematically impossible to satisfy all criteria for fairness simultaneously.[32] AI/ML thus poses a risk of learning and regurgitating biases in its training data, but no agency examined in the ACUS report has

---

22 *Id.* at 71–74.
23 *Id.*
24 *Id.* at 71.
25 *Id.*
26 *Id.* at 75–78.
27 *Id.*
28 *Id.* at 28.
29 *Id.*
30 *Id.* at 79.
31 *Id.* at 80.
32 *See* Virginia Foggo et al., *Algorithms and Fairness*, 17 OHIO ST. TECH. L.J. 123, 143–45 (2021). For example, an algorithm attempting to predict whether a job candidate is qualified across two groups with different rates of qualification cannot meet more than one of the following fairness criteria: both groups have the same chance of being accepted ("demographic parity"), qualified members of both groups have the same chance of being accepted ("equalized odds"), and accepted members of both groups have the same chance of being qualified ("predictive parity"). *Id.*; ENGSTROM ET AL., *supra* note 4, at 79.

established systematic protocols for assessing the risk or degree of bias in AI tools.[33]

Fourth, agencies adopting AI may endanger hearing rights and due process if AI fails to encompass normative concerns or pushes human decision makers out of the hearing process.[34] AI promises to improve the accuracy and efficiency of adjudicatory decisions, but it does so by de-emphasizing or even removing a human decision maker.[35] An AI approach to hearing rights may supplant the discretion of humans to make value judgments, either laying bare the requisite normative tradeoffs or obfuscating them behind a veneer of algorithmic objectivity.[36] Where AI fails to encapsulate these tradeoffs effectively, it may erode decision quality and public confidence in the agency, obviating any supposed efficiency gains. AI tools also blur the line between rulemaking and adjudication under the APA.[37] As systems grow more effective and adjudicatory decisions grow more reliant upon them, AI tools may start to function like legislative rules with a binding effect on regulated parties, which would trigger requirements for notice and comment. However, the precise moment of transition from adjudicatory tool to rule may be difficult to spot.

Fifth, the report voices concerns with regulatory targets gaming AI systems by learning how to navigate or reverse-engineer them.[38] The risks are compounded in situations where only sophisticated actors would be capable of such adversarial learning, creating a competitive advantage over other regulatory targets.[39] The buy-or-build decision again poses concerns. Where an agency contracts out for an AI tool, the contractor may seek to leverage or monetize its relationship with and understanding of the tool in other business relationships.[40]

Sixth, agencies contracting out for AI tools to circumvent their lack of technical expertise and capacity pose additional risks of conflicts of interest.[41] Although embedding technical expertise at an agency may better automate the agency's workflows, allow additional responsiveness to changing circumstances, and limit the leakage of information regarding how algorithmic tools work, agencies will be tempted to buy external tools in hopes that those AI systems are cheaper and more effective than developing internal expertise.[42] However, when an agency needs AI tools that can adapt frequently, contracting

---

33 ENGSTROM ET AL., *supra* note 4, at 80–81.
34 *Id.* at 82–85.
35 *Id.*
36 *Id.* at 83.
37 *Id.* at 84.
38 *Id.* at 86–87.
39 *Id*.
40 *Id*.
41 *Id.* at 88–90.
42 *Id.* at 88–89.

may be an inefficient solution, on top of being a potential source of conflicts of interest. A third choice for agencies seeking external expertise is to engage in noncommercial collaboration, for example with universities or by sponsoring public competitions with prize money.[43]

II. HARD LOOK REVIEW

Private parties may challenge agency action on substantive and procedural grounds.[44] Procedural challenges attack the amount or types of process that an agency followed in taking a particular action, without considering whether the action was ultimately correct.[45] Substantive challenges provide that remaining element, attacking agency action based on its soundness.[46] Substantive challenges are not without critics, who argue that permitting courts into the weeds of core agency functions violates the separation of powers.[47] Substantive review may also imperil agency expertise by permitting lay judges to review agency decisions for technical correctness.[48] Substantive review can increase the cost of agency action by burdening the agency with a requirement to create and maintain documentation, possibly compromising the agency's ability to respond flexibly to dynamic regulatory environments.[49] This can in turn ossify agency action and limit agencies' capacity to innovate and evolve.[50] These concerns underlie a fear that substantive review will actually decrease transparency into the motivations for agency action as agencies scrub official sources of their actual motivations and craft sanitized records for judicial review around more

---

43 *Id.* at 90.

44 *See* 5 U.S.C. § 704 (courts can review all "final agency action").

45 *See* 5 U.S.C. § 553 (describing the procedural requirement for notice and comment review); 5 U.S.C. §§ 556–557 (adding additional procedural requirements for formal rulemaking).

46 *See* 5 U.S.C. § 704 (final agency action is subject to judicial review).

47 Ethyl Corp. v. EPA, 541 F.2d 1, 38 (D.C. Cir. 1976) (Bazelon, C.J., concurring) ("[I]n cases of great technological complexity, the best way for courts to guard against unreasonable or erroneous administrative decisions is not for the judges themselves to scrutinize the technical merits of each decision. Rather, it is to establish a decision-making process that assures a reasoned decision that can be held up to the scrutiny of the scientific community and the public." (quoting Int'l Harvester Co. v. Ruckelshaus, 478 F.2d 615, 652 (D.C. Cir. 1973) (Bazelon, C.J., concurring)).

48 *See* Louis J. Virelli III, *Deconstructing Arbitrary and Capricious Review*, 92 N.C. L. REV. 721, 765–66 (2014).

49 *Id.* at 773–75.

50 *See* Jason Webb Yackee & Susan Webb Yackee, *Testing the Ossification Thesis: An Empirical Examination of Federal Regulatory Volume and Speed, 1950-1990*, 80 GEO. WASH. L. REV. 1414, 1427–28 (2012); *see also* Richard J. Pierce, Jr., *Rulemaking Ossification Is Real: A Response to Testing the Ossification Thesis*, 80 GEO. WASH. L. REV. 1493, 1499 (2012) (suggesting ossification of rulemaking preceded the 1970s).

defensible justifications.[51]

On the other hand, the existence of substantive challenges puts agencies on notice that they must produce credible reasons for a decision, leading, at least in theory, to better decision-making by requiring agencies to justify and vet their decisions.[52] Substantive review may also help to prevent agency capture and other forms of bad faith agency action.[53] The desirability of agency action motivated by political goals is debatable, but by forcing transparency through substantive review, agencies must at a minimum admit when politics motivate their decisions.[54] Substantive review also permits courts to provide additional checks on agency action, which may be valuable at a time when legislative checks like the nondelegation doctrine and legislative veto are in disrepute.[55]

A fierce debate over the merits of substantive review took place within the D.C. Circuit in the 1960's between Judge Harold Leventhal, who made the above arguments in favor of substantive review, and Chief Judge David Bazelon, who asserted that procedural review alone is a sufficient check on agency action.[56] The Supreme Court settled the debate in *Vermont Yankee Nuclear Power Corp. v. Natural Resources Defense Council, Inc*., unanimously rejecting Judge Bazelon's position for procedural-only review.[57] But that conclusion is just the tip of the iceberg. Once we accept that courts are permitted to probe the substance of agency action, we must then determine how much of their own judgment they are permitted to substitute for agency action: how "hard" a look should courts take at the substance of agency action? To this question, addressed below, the Bazelon-Leventhal debate is eminently relevant.

A. The Traditional Account and *State Farm*

Section 706(2)(A) of the APA requires courts to "hold unlawful and set aside" agency action that is "arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law."[58] Agency action is arbitrary and capricious if the agency fails to consider relevant factors in its decision, considers impermissible

---

51 Lukas Gemar, *Rejecting Two-Faced Explanations by Agencies*, THE REGULATORY REVIEW (Nov. 25, 2021), https://www.theregreview.org/2021/11/25/gemar-rejecting-two-faced-explanations-by-agencies/.

52 Virelli, *supra* note 49, at 743.

53 *Id.* at 775.

54 *Id.* at 771–72.

55 *See* Richard W. Murphy, *The Limits of Legislative Control over the "Hard-Look,"* 56 ADMIN. L. REV. 1125, 1131–35 (2004).

56 *Compare* Ethyl Corp. v. Env't Prot. Agency, 541 F.2d 1, 66–68 (D.C. Cir. 1976) (Bazelon, J., concurring), *with id*. at 68–69 (Leventhal, J., concurring).

57 Vermont Yankee Nuclear Power Corp. v. Nat. Res. Def. Council, Inc., 435 U.S. 519, 542–43 (1978).

58 5 U.S.C. § 706(2)(A).

factors, or makes a clear error in judgment; otherwise, the action is rational.[59] Courts can review all "final" agency action unless one of two narrow exceptions is met: a statute conveys clear legislative intent to preclude judicial review, or there is no meaningful law or standard to apply in judicial review.[60] In practice, courts tend to find agency action arbitrary where the agency failed to give reasons for its decision, did not produce sufficient reasons for the decision, or provided impermissible reasons for the decision.[61]

Since the enactment of the APA, courts have struggled to determine how deep into the details of agency reason-giving they must reach to determine whether an action is arbitrary and capricious.[62] The Supreme Court has elaborated on the requirements of rational basis review, most famously in three cases: *Citizens to Preserve Overton Park, Inc. v. Volpe* ("*Overton Park*"), *Motor Vehicle Manufacturers Ass'n of U.S., Inc. v. State Farm Mutual Automobile Insurance Co.* ("*State Farm*"), and *FCC v. Fox Television Stations, Inc.* ("*FCC v. Fox*").[63] In *Overton Park*, the Court reviewed a determination that the Secretary of Transportation did not violate the Department of Transportation Act or Federal Highway Act when he authorized the construction of a highway through a public park because no feasible alternative route existed.[64] The Court held that the

---

59 Citizens to Pres. Overton Park, Inc. v. Volpe, 401 U.S. 402, 415 (1971) ("Certainly, the Secretary's decision is entitled to a presumption of regularity. But that presumption is not to shield his action from *a thorough, probing, in-depth review*." (emphasis added) (citations omitted)); White Stallion Energy Ctr., LLC v. E.P.A., 748 F.3d 1222, 1233 (D.C. Cir. 2014) ("The 'arbitrary and capricious' standard deems the agency action presumptively valid provided the action meets a minimum rationality standard." (quoting Sierra Club v. E.P.A., 353 F.3d 976, 978 (D.C. Cir. 2004))), *rev'd sub nom.* Michigan v. E.P.A., 576 U.S. 743 (2015).

60 *Citizens to Pres. Overton Park, Inc.*, 401 U.S. at 410.

61 Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 50 (1983) ("It is well established that an agency's action must be upheld, if at all, on the basis articulated by the agency itself."); Pac. Coast Fed'n of Fishermen's Ass'ns v. U.S. Bureau of Reclamation, 426 F.3d 1082, 1091 (9th Cir. 2005) ("It is a basic principle of administrative law that the agency must articulate the reason or reasons for its decision."); Deukmejian v. Nuclear Regul. Comm'n, 751 F.2d 1287, 1326 n.244 (D.C. Cir. 1984) ("Courts disregard post hoc rationalizations of an agency's position on preexisting records." (emphasis omitted)).

62 *Compare* Ethyl Corp. v. EPA, 541 F.2d 1, 67 (D.C. Cir. 1976) (Bazelon, J., concurring) (arguing for a procedural common law in place of substantive review of agency action), *with id.* at 68–69 (Leventhal, J., concurring) (supporting hard look review); *see also* Nat. Res. Def. Council, Inc. v. U.S. Nuclear Regul. Comm'n, 547 F.2d 633, 654 (D.C. Cir. 1976), *rev'd sub nom.* Vt. Yankee Nuclear Power Corp. v. Nat. Res. Def. Council, Inc., 435 U.S. 519 (1978), *cert. granted, vacated sub nom.* Balt. Gas & Elec. Co. v. Nat. Res. Def. Council, Inc., 435 U.S. 964 (1978).

63 *See generally* Citizens to Pres. Overton Park, Inc. v. Volpe, 401 U.S. 402 (1971); Motor Vehicle Mfrs. Ass'n U.S., Inc. v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29 (1983); FCC v. Fox Television Stations, Inc., 556 U.S. 502 (2009).

64 *Overton Park*, 401 U.S. at 404–06.

lower courts erred in making their determinations by relying on documents developed for litigation, instead of an administrative record.[65] In laying out the initial test for whether an agency action is arbitrary and capricious, Justice Marshall admitted that it might be impossible for the district court to make a determination from the existing administrative record, in a tacit acknowledgement that agencies had not previously had notice that courts would scour such records to determine if agency action is rational.[66]

In *State Farm*, the Court reviewed the Secretary of Transportation's decision to revoke the National Highway and Traffic Safety Administration's (NHTSA) safety standard requiring passive restraints in vehicles promulgated under authority from the National Traffic and Motor Safety Act.[67] Writing for the Court, Justice White added gloss to the *Overton Park* test by centering the inquiry on whether the agency (1) relied on the wrong factors, (2) failed to consider an obvious alternative, or (3) gave an explanation that runs counter to the evidence or cannot be a product of the agency's expertise.[68] He reached deep into NHTSA's decision, determining that the agency gave no reasons for failing to pursue an obvious alternative (pursuing regulation requiring only airbags), and that the agency wrongfully extrapolated from research in deciding to pursue regulation requiring only detachable seatbelts.[69]

Finally, in *FCC v. Fox*, the FCC had previously changed its stance on exceptions to its obscenity rules in a prior adjudication, and it applied the new position in the instant case.[70] Justice Scalia's majority concluded that the agency's change in policy was not arbitrary and capricious because the reasons the agency gave for its decision "were entirely rational."[71] Unlike the majorities in *Overton Park* and *State Farm*, the Court addressed the quality of the agency's proffered reasons, instead of their relation to the agency's ultimate policy determination.[72] This additional deference might signal a retreat from the presumptive hardness of *State Farm* rationality review.[73]

B. The Empirical Account and *Baltimore Gas*

The progression of hard look review recounted above describes different approaches and levels of stringency through which the Court has sanctioned this

---

65 *Id.* at 419–20.
66 *Id.* at 420.
67 *State Farm*, 463 U.S. at 33–34.
68 *Id.* at 43.
69 *Id.* at 50–51, 53–54.
70 FCC v. Fox Television Stations Inc., 556 U.S. 502, 508–10, 512.
71 *Id.* at 517.
72 *See Id.* at 529–30; *see also* Virelli, *supra* note 49, at 732.
73 *See* Virelli, *supra* note 49, at 732.

doctrine's application. The *Overton Park* Court policed (and progenerated) agency record-producing, the *State Farm* Court reviewed agency rationality, and the *FCC v. Fox* Court addressed only agency reason-giving.[74] Courts often read (and rotely cite) *State Farm* as establishing a regime of "hard" hard look review.[75] But Professor Louis Virelli argues that this line of cases shows the Court deconstructing arbitrary and capricious review to tailor judicial deference to the particular policymaking activity the agency is engaged in.[76] *FCC v. Fox*, with its emphasis on agency reason-giving, signaled that the Court, in practice, employs a more deferential approach, especially where agency action implicates the agency's considered expertise.[77]

The Court crystallized this deferential position in *Baltimore Gas & Electric Co. v. Natural Resources Defense Council, Inc.*, which it handed down in the same term as *State Farm*.[78] In this case, Justice O'Connor upheld a determination by the Nuclear Regulatory Commission that the National Environmental Policy Act ("NEPA") permitted the agency to assume that permanent storage of certain nuclear waste had no significant environmental impact.[79] Though the D.C. Circuit (spearheaded by Judge Bazelon) had struck down the decision as arbitrary for failing to consider the significant environmental risks of its assumption policy, a unanimous Court reversed.[80] Justice O'Connor wrote that where an agency is making predictions within its area of "special expertise," "as opposed to simple findings of fact, a reviewing court must generally be at its most deferential."[81] Professors Gerson and Vermeule argue that *Baltimore Gas*, more than *State Farm*, articulates the throughline of the Court's jurisprudence on judicial deference because the holding tethers deference to the expertise the agency leverages in taking action.[82]

C. Squaring the Cases

Which account of rationality review should we prefer? *Baltimore Gas* and *State Farm* came down in the same term, and their approaches to the

---

74 Citizens to Pres. Overton Park v. Vlope, 401 U.S. 402, 415, 419–20 (1971); *State Farm*, 463 U.S. at 31, 42–43; FCC v. Fox, 556 U.S. at 552–54; *see also* Virelli, *supra* note 49, at 728–33.

75 Jacob Gersen & Adrian Vermeule, *Thin Rationality Review*, 114 MICH. L. REV. 1355, 1358–60 (2016).

76 Virelli, *supra* note 49, at 729–33.

77 *Id*. at 750–52.

78 *See* Balt. Gas & Elec. Co. v. NRDC, 462 U.S. 87, 88–90, 103 (1983).

79 *Id*. at 106–07.

80 *Id*. at 95, 108.

81 *Id*. at 103.

82 Gersen & Vermeule, *supra* note 76, at 11355, 359–60.

arbitrariness inquiry are difficult to square.[83] *Baltimore Gas* posits the common-sense position that courts should adjust the level of deference to an agency action to the extent to which that action implicates the agency's substantive expertise.[84] But *State Farm* dug deep into the agency's decision despite its acknowledgement of the extent of the agency's expertise in taking the instant action.[85] In resolving the tension, we begin with an empirical review of the other cases on agency action the Court has taken. The Court's record has been empirically unequivocal in its preference for deference.[86] In its sixty-four cases involving an arbitrary and capriciousness determination on agency action between 1983 and 2014 (excluding *State Farm*), only eight agency actions were found to be arbitrary and capricious.[87] A recent example of the Court's deferential bent at work is *EPA v. EME Homer City Generation, L.P*.[88] The D.C. Circuit had held that the Clean Air Act required the EPA to apportion obligations for reducing windborne interstate air pollution among states based on each state's physical contribution, though the agency had previously allocated obligations based on a cost-benefit analysis of the impact of reducing particular emissions.[89] The Court, however, held that statutory silence on the dispute permitted the agency to act however it saw fit.[90] A similar example of deference is *Utility Air Regulatory Group v. EPA*, penned by Justice Scalia, in which the Court held that the EPA was free to interpret the occurrences of the term "air pollutant" in the Clean Air Act differently, running directly counter to the canon of consistent usage (arguably surprising in an opinion by Justice Scalia).[91]

---

83 *See generally Balt. Gas & Elec. Co.*, 462 U.S. 87 (1983); Motor Vehicle Mfrs. Ass'n U.S., Inc. v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29 (1983).

84 *Balt. Gas & Elec. Co.*, 462 U.S. at 103.

85 *State Farm*, 463 U.S. at 53–54.

86 Gersen & Vermeule, *supra* note 76, at 1407.

87 *Id.* These cases included six challenges to enforcement orders, permit decisions, and notice and comment rules levied against the Environmental Protection Agency; two Board of Immigration proceedings; five challenges to rate-setting, enforcement actions, and informal rulemakings at the Federal Communications Commission; and fourteen challenges to the Department of Health and Human Services (HHS). *Id.* at 1364.

88 *See* EPA v. EME Homer City Generation, L.P., 572 U.S. 489, 524 (2014).

89 EME Homer City Generation, L.P. v. EPA, 696 F.3d 7, 26–27 (D.C. Cir. 2012), *rev'd*, 572 U.S. 489 (2014).

90 EPA v. EME Homer City Generation, L.P., 572 U.S. at 514–15.

91 Util. Air Regul. Grp. v. EPA, 573 U.S. 302, 319–20 (2014) (finding against the agency, which argued that its hands were tied in interpretation of statutory terms by its other interpretations of the same term throughout the statute). It would be remiss, in a discussion of the current law on judicial review of agency action, not to mention *West Virginia v. EPA*, in which the Court limited the scope of the EPA's power in regulating emissions under the Clean Air Act. 142 S. Ct. 2587, 2616 (2022). Though the D.C. Circuit had found the agency's action arbitrary and capricious below, rationality review did not make an appearance in the Court's opinion, which focused on the major questions doctrine in *Chevron* analysis. *Id.* at 2610.

At the opposite end of the spectrum and a relative outlier, the recent *Judulang v. Holder* provides an example of the Court's arbitrary and capriciousness review at its "hardest."[92] But the Court's deference even in finding the action irrational is still so great that the case serves as an exception to prove the rule.[93] The Court considered a challenge to the Board of Immigration Appeals' "comparable grounds" rule, which permitted appeals to the Attorney General for relief from deportation only when the charged deportation ground has a close analogue to a grounds for exclusion in the Immigration and Nationality Act.[94] The Court held the policy to be arbitrary and capricious, reasoning that "[w]hen an administrative agency sets policy, it must provide a reasoned explanation for its action."[95] Here, "[r]ather than considering factors that might be thought germane to the deportation decision, that policy hinges § 212(c) eligibility on an irrelevant comparison between statutory provisions."[96] This version of "hard look" review is more modest than was applied in *State Farm*, where the Court required the agency to weigh the factors Congress set out for the agency by statute, because the Court indicated it would have deferred had the agency even tangentially supported its policy with factors "tied, even if loosely, to the purposes of the immigration laws or the appropriate operation of the immigration system."[97]

In addition, a more circumspect reading of *State Farm* shows that it is consistent with the *Baltimore Gas* approach. *State Farm* based its finding of arbitrariness on three separate grounds: NHTSA's failure to consider an airbags-only standard, as opposed to allowing the manufacturer to choose between passive seat belts and airbags; the agency's inadequate consideration of automatic seatbelts before rejecting their requirement; and the agency's failure to consider nondetachable seatbelts separately from other belts as part of its safety standard.[98] First, the Court predicated the arbitrariness of the agency's failure to consider mandating airbags alone on the agency's failure to provide any reasons at all for that decision.[99] The Court was expressing concern with the agency's "first-order" reason giving, its ability to give any reasons at all (as opposed to "second-order" reason giving, providing reasons that can persuade a

---

92 *See* Judulang v. Holder, 565 U.S. 42, 45 (2011); Gersen & Vermeule, *supra* note 76, at 1362–63.

93 Gersen & Vermeule, *supra* note 76, at 1363.

94 *Judulang*, 565 U.S. at 59–63; *see In re* Blake, 23 I. & N. Dec. 722, 728 (2005); *See generally* 8 U.S.C. § 1182.

95 *Judulang*, 565 U.S. at 45.

96 *Id.* at 55.

97 Motor Vehicle Mfrs. Ass'n U.S., Inc. v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 43 (1983); *Judulang*, 565 U.S. at 55.

98 *State Farm*, 463 U.S. at 46–56.

99 *Id.* at 50.

court that the agency is right).[100] *Baltimore Gas* and *Judulang* support this form of arbitrariness review, since even soft look deference to agency expertise requires reason giving of the agency.[101] Second, the Court held NHTSA's decision to reverse course on the automatic seatbelt requirement was arbitrary based on second-order reason giving, determining that there was no "rational connection between the facts found and the choice made."[102] This determination poses the most direct challenge to an expertise-sensitive, broadly deferential standard. And yet, on this issue the Court still couched its holding in terms of the expertise-based, adjustable deference of *Baltimore Gas*, requiring the agency to "bring its expertise to bear on the question" to be entitled to deference.[103] Based on the agency's findings that a significant portion of drivers used seatbelts at least some of the time, and that the "inertia" of a restraint currently in place tended to be significant in individuals' determinations to use restraints, the Court, per Justice White, thought it was obvious that a restraint that continues to function automatically unless disconnected would increase seatbelt use, contrary to the agency's conclusion otherwise.[104] One reading of *State Farm* is that the Court saw the agency's determination of whether people will use detachable automatic seatbelts as implicating a small enough helping of the agency's expertise that the Court was permitted to wade in and correct what it saw, rightly or wrongly, as an obviously logically flawed position.[105] Read this way, the Court is suggesting that the decision to use hard look review involves gauging the agency's demonstrated expertise not only through an assessment of the problem the agency is acting to solve, but through the solution that it provides. Where the agency obviously does not bring its substantive expertise to bear in an action (such as by making a glaring logical error in its reasoning), even though action in good faith would obviously require the use of such expertise, a court is permitted to consider the agency's obvious lack of credibility in its ultimate

---

100 Virelli, *supra* note 49, at 744; *State Farm*, 463 U.S. at 50.

101 Virelli, *supra* note 49, at 744.

102 *State Farm*, 463 U.S. at 52 (quoting Burlington Truck Lines, Inc. v. United States, 371 U.S. 156, 168 (1962)).

103 *State Farm*, 463 U.S. at 54.

104 *Id.*

105 The Court's intrusion here is reminiscent of a phenomenon noted in patent law cases by Herbert F. Schwartz in which judges are historically more likely to find mechanical inventions to be "obvious" than inventions in technical fields with higher knowledge barriers to entry, in spite of expert examiners' findings to the contrary. *See generally* John Richards et al., *Panel I: KSR v. Teleflex: The Non-Obviousness Requirement of Patentability*, 17 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 875, 889 (2007) (in which Mr. Schwartz outlined parts of his views on patent obviousness and the meaning of "ordinary skill in the art"). This trend, taught to me under the name "real men know mechanicals," be they obvious inventions or the usage patterns of seatbelts, would indicate that the Court underestimated the agency's expertise in extrapolating from the studies here. Thanks to Professor Rochelle Dreyfuss of NYU for relating this story.

determination on how much deference is owed. This interpretation is not only compatible with *Baltimore Gas*, but helps to develop it into a more workable standard capable of addressing agencies acting in bad faith in situations where the requisite expertise would otherwise prevent courts from questioning the agency's action.[106]

What, then, explains the hallowed status of *State Farm* and hard look review, if such a reading of the case runs so contrary to the majority of the Court's holdings? In part, *State Farm*'s authorization of expansive review is popular with the courts of appeals.[107] One study of published appellate reviews of EPA and NLRB decisions between 1996 to 2006 found that those agencies won only 64 percent of arbitrariness challenges.[108] This figure is to be compared against the 87 percent agency success rate in arbitrariness challenges before the Court that Gerson and Vermeule found[109] and the 69 percent agency success rate in litigation overall found in a meta-analysis of eleven studies.[110] Though not enough to show a clear preference for hard look review in the courts of appeals, the relative disinclination to defer to agencies indicates that courts may be drawn to a version of rationality review that grants them more power to review.[111] The D.C. Circuit, for example, has cited *State Farm* in 830 opinions to date, while it has cited *Baltimore Gas* in 76 opinions, and it has primarily cabined use of the case to the specific statute it discussed, NEPA.[112]

---

106 One might wonder, however, if an inquiry that permits a court to assess the quality of the reasons an agency gives in determining whether to defer to the agency or supplant the agency's judgment with its own, is any different from hard look review in the first place. This echoes the common criticism of the *Chevron* doctrine that the whole inquiry ultimately collapses into a reasonableness test, where the added steps provide little clarity despite providing grounds for disagreement and reversal.

107 Gersen & Vermeule, *supra* note 76, at 1364–65.

108 Thomas J. Miles & Cass R. Sunstein, *The Real World of Arbitrariness Review*, 75 U. CHI. L. REV. 761, 766, 813 (2008).

109 Gersen & Vermeule, *supra* note 76, at 1364, 1407; *Id.* at 777–78 (illustrating the grounds for doubting the aptness of this comparison; nearly 85% of Miles' and Sunstein's cases are NLRB cases, which they concede is an outlier agency both in approach to adjudications, using them to announce rules for general application, and its ideological candor, possessing a "liberal" bent); *Id.* at 778–79 (illustrating that as for the EPA, decisions survived arbitrariness challenges almost 80% of the time, and "major" determinations survived challenges nearly 75% of the time).

110 David Zaring, *Reasonable Agencies*, 96 VA. L. REV. 135, 169–70 (2010) (illustrating that the overall agency win rate in litigation was fairly stable across all considered standards of review (e.g., *Chevron*, hard look, and substantial evidence)). Again, there are concerns with making this comparison directly. Agencies must often seek the Solicitor General's approval to petition the Court for certiorari, which can prevent weaker cases from reaching the Court. Also, agencies may implicitly adapt the scope and defensibility of their actions to the stringency of judicial review to stabilize their litigation win rates at any level of review).

111 *See* Gersen & Vermeule, *supra* note 76, at 1364–67.

112 Citation counts are a result of an independent Westlaw search. *See, e.g.*, WildEarth

Another reason for *State Farm*'s enduring significance is that the case is now more often regurgitated as a symbol than it is read for its holdings, much like *Marbury v. Madison*.[113] In *State Farm*, the Court was clear that the agency's discretion regarding permissible extrapolations from research on seatbelt use was "precisely the type of issue which rests within the expertise of NHTSA, and upon which a reviewing court must be most hesitant to intrude."[114] The decision also makes clear that courts cannot require agencies to consider any policy alternative that the court prefers.[115] *FCC v. Fox* agreed, holding that agencies are not generally required to engage in comparative policy evaluation.[116] And *State Farm* certainly does not prevent an agency from acting where the agency has good reasons for acting in general, despite being unable to justify the particular chosen action (as where the agency is choosing between two equally good alternatives).[117] Flatly citing *State Farm* for the proposition that courts can engage in hard look review of agency action is in tension not only with the Court's deferential administrative law doctrines[118] and history of empirically demonstrable deference,[119] but with the *State Farm* decision itself.

Guardians v. Jewell, 738 F.3d 298 (D.C. Cir. 2013).

113 *See* Marbury v. Madison, 5 U.S. 137, 179–180 (1803). While the reasoning of the case was narrow, the case now stands for a broad statement of judicial supremacy independent of its facts and holding. *See, e.g.*, Zivotofsky v. Clinton, 132 S. Ct. 1421, 1427–28 (2012) (under *Marbury*, "when an Act of Congress is alleged to conflict with the Constitution, '[i]t is emphatically the province and duty of the judicial department to say what the law is.'" (citation omitted) (quoting Marbury v. Madison, 5 U.S. 137 (1803))).

114 Motor Vehicle Mfrs. Ass'n U.S., Inc. v. State Farm Mut. Auto. Ins. Co., 463 U.S. 29, 53 (1983) (deciding not to intrude in exactly that determination regardless).

115 *Id.* at 51.

116 *See State Farm*, 463 U.S. at 51 (quoting Vt. Yankee Nuclear Power Corp. v. Nat. Res. Def. Council Inc., 435 U.S. 519, 551 (1978)) ("Nor do we broadly require an agency to consider all policy alternatives in reaching decision. It is true that rulemaking 'cannot be found wanting simply because the agency failed to include every alternative device and thought conceivable by the mind of man . . . regardless of how uncommon or unknown that alternative may have been . . . .'"); *see also* Pension Benefit Guar. Corp. v. LTV Corp., 496 U.S. 633, 646 (1990) ("If agency action may be disturbed whenever a reviewing court is able to point to an arguably relevant statutory policy that was not explicitly considered, then a very large number of agency decisions might be open to judicial invalidation."); FCC v. Fox Television Stations, Inc., 556 U.S. 502, 514 (2009).

117 *See* Gersen & Vermeule, *supra* note 76, at 1387 ("[T]he nature of the rational connection [between facts and choices] . . . arises because the agency has good reason to decide, even if it lacks good reason for the decision. In this sense, the *State Farm* test of rational connection is more capacious, more forgiving, and less demanding than is conventionally understood.").

118 *See, e.g.*, Chevron, U.S.A., Inc. v. Nat. Res. Def. Council, Inc., 467 U.S. 837, 865 (1984).

119 *See, e.g.*, Balt. Gas & Elec. Co. v. NRDC, 462 U.S. 87, 103 (1983); Util. Air Regul. Grp. v. EPA, 573 U.S. 302, 321 (2014).

III. THE DEFERENCE INQUIRY FOR AGENCY ACTION BASED ON AI

Algorithmic tools can function as a focal point for judicial review.[120] But understanding the substance of an AI as part of reviewing agency action will be a daunting challenge for courts, one which may convince them to shake off the rote presumption of *State Farm*-inspired hard look review. Too much deference to an agency's use of AI may allow the agency to wield unchecked a tool that the agency itself does not understand. Too little deference, and the court may find itself supplanting the judgments of algorithmic experts with its own. This Scylla and Charybdis are shared by all judicial review of agency action, but there is additional cause for wariness when the agency action is based on AI because, in the context of such an action, both the court and the agency are likely to be even further out of their areas of expertise than in the context of more standard agency actions.

This paper's main contribution is a framework by which courts can gauge the appropriate level of deference for agency action based on AI. The choice between harder or softer look review of a given AI-based action should, of course, use the *Baltimore Gas* approach: does the agency's action implicate its substantive expertise? Additionally, the policy concerns of substantive review expressed in the Bazelon-Leventhal debate inform the discussion of the hardness of review.[121] Moreover, because *Baltimore Gas* invites a case-by-case determination of the expertise an agency brings to bear in any given action, there is no one-size-fits-all answer for every agency action or every AI-based tool used. A realistic inquiry requires courts to understand the AI and its role in the agency's action and to identify and assess the characteristics of the AI which entitle its impact on the action to more or less deference. The relevant factors

---

120 *See generally* DILLON REISMAN ET AL., ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY (2018) (arguing that a tendency to center judicial review on AI tools can be detrimental, where those tools perform functions that are typically walled off from judicial review, such as an agency's decision to institute an enforcement proceeding); *see* Heckler v. Chaney, 470 U.S. 821, 832–33 (1985) (holding that agency decisions not to undertake enforcement actions are presumptively unreviewable); ENGSTROM ET AL., *supra* note 4, at 77 (suggesting that Congress relax the *Heckler* presumption against reviewability to permit judicial review of algorithmic enforcement tools, in conjunction with or in place of ex ante review through notice-and-comment, because algorithmic tools can function more like legislative rules than an exercise of agency discretion).

121 Remember that the policy considerations of the Bazelon-Leventhal debate discussed *supra* are on the one hand (counseling in favor of soft look review) separation of powers concerns, the relative inexpertise of courts as compared to agencies, and the lost agency flexibility and ossification that result from an increased burden on the agency of hard look review; and on the other (counseling in favor of hard look review) forcing agencies to make more defensible (and thus theoretically better) choices, preventing agency capture, promoting transparency, and supplementing existing checks on agency power.

include (1) whether the agency contracts for the AI or develops it in-house, (2) whether the agency action is explainable, (3) how the agency procured its data, (4) whether there is a human in the loop, (5) what use case the agency is fulfilling with the AI, and (6) the technical details of the agency's implementation of the AI. This set of factors influences the deference decision under the frameworks both of *Baltimore Gas* and of the Bazelon-Leventhal debate. I argue that these factors are comprehensive, setting forth a complete analysis of deference due to agency action based on AI.

A. The "Buy or Build" Decision

Agencies develop just over half of their AI tools in house, with the remaining use cases split between commercial contractors (the next most common source) and noncommercial collaborations (a distant third).[122] The government is proscribed from contracting out performance of "inherently governmental functions," and agency determinations of whether contracting is permissible for a given function are subject to review by the Office of Management and Budget.[123] Commentators have noted that the government's dependency on contractors and lack of information when participating in the market for AI tools can exacerbate the challenges of algorithmic governance.[124] AI procurement may be of custom AI solutions or of commercial off-the-shelf ("COTS") tools and may be performed with varying degrees of collaboration between the agency's internal teams and the contractor.[125] Conventional procurement regulations address COTS tools pragmatically, attempting to emulate a transaction on the commercial market.[126] After all, if a product is good enough for the private sector, why would it not also meet the government's needs? Professor David Rubenstein, however, argues that these assumptions hold less weight in the context of AI procurement, where the design and development of commercially available plug-and-play AI solutions is largely unregulated.[127] The use of COTS solutions also limits an agency's ability to configure tools to its use cases (for example, by training the AI on its own data), or at a minimum

---

122 ENGSTROM ET AL., *supra* note 4, at 88.

123 *See* 48 C.F.R. § 7.503(a)–(b) (2021); *see also* 48 C.F.R. § 37.104(c)(2) (2021) ("Each contract arrangement must be judged in the light of its own facts and circumstances, the key question always being: Will the Government exercise relatively continuous supervision and control over the contractor personnel performing the contract?").

124 David S. Rubenstein, *Acquiring Ethical AI*, 73 FLA. L. REV. 747, 754 (2021); ENGSTROM ET AL., *supra* note 4, at 86; *see* Cary Coglianese & Erik Lampmann, *Contracting for Algorithmic Accountability*, 6 AM. U. ADMIN. L. REV. 175, 198 (2021).

125 Rubenstein, *supra* note 125, at 813–14.

126 *Id.*

127 *Id.*

adds the costs and challenges of long-term support and maintenance.[128] Custom AI solutions, on the other hand, may be better suited to the government's particularized needs, but customized procurement is, again, uniquely difficult in the case of AI.[129] The development process is complex, unpredictable, and laden with trial and error, preventing the agency from fully fixing all design and technical requirements at time of purchase.[130] Thus, even when agencies decide to purchase AI externally, they still must develop substantial in-house expertise to supervise long-standing iterative development contracts with the private sector.[131]

In-house development can also take several forms. Tools can be developed entirely internally, or through collaboration with the public, through universities, NGOs, or agency-sponsored competitions.[132] The ACUS report notes that in-house technical expertise, if not complete in-house development, leads to tools which "are better tailored to complex governance tasks and more likely to be designed and implemented in lawful, policy-compliant, and accountable ways."[133] In-house development can also help to produce flexible tools since the agency that makes a tool is better equipped to foresee necessary changes and adapt it on the fly.[134] Finally, in-house development can limit leaks of a tool's technical and operational details, which makes attacks and gaming more difficult while also securing the privacy and security of the data set.[135]

Professor Cary Coglianese and Erik Lampmann observe that when the government contracts for AI, there is tension between "vendors' legitimate trade secrets and the necessity of providing members of the public some form of visibility into automated government decisionmaking."[136] They also note data privacy and security risks and associated reputational, financial, and technical harms in contracting for third-party AI tools.[137] The public may also lose opportunities to participate in the design, development, and evaluation of algorithms, resulting in neglecting an essential source of feedback in important

---

128 *Id.* at 814.

129 *Id.* at 814–17.

130 *Id.* at 814.

131 NAT'L SEC. COMM'N ON ARTIFICIAL INTEL., FINAL REPORT 123 (2021) (noting that agencies that "rely solely on contractors for digital expertise will become incapable of understanding the underlying technology well enough to make successful acquisition decision independent of contractors").

132 ENGSTROM ET AL., *supra* note 4, at 7.

133 *Id.*

134 *Id.* at 88.

135 *Id.* at 88.

136 Coglianese & Lampmann, *supra* note 125, at 184. *See* Christopher J. Morten, *Publicizing Corporate Secrets*, 171 U. PA. L. REV. (forthcoming 2023) (arguing that corporate trade secrets can be publicized for public good).

137 Coglianese & Lampmann, *supra* note 125, at 189–90.

decisions (such as how to balance tradeoffs between accuracy and equity).[138] Coglianese and Lampmann suggest requiring algorithmic impact statements or risk management plans to identify the purpose, challenges, and concerns of AI uses, as well as regular audits of the technology and security practices.[139]

On the other hand, AI developed internally at agencies is not above criticism. The federal government is faced with a shortage of AI talent, and it would seem foolish to establish so strong a preference for in-house development that an agency is expected to create its own algorithmic tools when it is fundamentally incapable of doing so.[140] Compared to private sector efficiency, in-house tools can be expensive.[141] These problems are compounded where an agency's leadership fails to prioritize technological innovation and internal capacity development through compensation caps and other limitations.[142]

The *Baltimore Gas* expertise inquiry weighs heavily in favor of granting more deference for in-house development. An agency which purchases an AI tool, whether COTS or custom, is leveraging its substantive expertise to a lesser extent than an agency which builds the same tool in house. Even if we concede that an agency must leverage its expertise to draft the requirements for the AI use case that the contractor will eventually implement (which is only possible with custom solutions, not with COTS), the agency would have to draft the same requirements as the first step of building the AI itself, implicating all the same expertise. From there, the two tools diverge, with the hypothetical in-house tool likely being the subject of closer collaboration between users and developers compared to the hypothetical purchased tool. Presumably, the agency will better understand the functioning and compliance of the in-house tool, not to mention the ACUS report's finding that the agency will generally make a better tool.[143] All these considerations point to additional expertise and, accordingly, additional deference.

The Bazelon-Leventhal factors counsel a similar structure for adjusting deference based on procurement, though not as definitively. Where an agency procures an AI tool, the agency is at risk of industry capture if a contractor

---

138 *Id*. at 194–95.

139 *Id*. at 192–93.

140 *See* Jody Freeman, *Extending Public Law Norms Through Privatization*, 116 HARV. L. REV. 1285, 1296 (2003) ("[P]rivatization is a means of improving productive efficiency: obtaining high-quality services at the lowest possible cost").

141 *See* PAUL R. VERKUIL, OUTSOURCING SOVEREIGNTY: WHY PRIVATIZATION OF GOVERNMENT FUNCTIONS THREATENS DEMOCRACY AND WHAT WE CAN DO ABOUT IT (2007); ENGSTROM ET AL., *supra* note 4, at 89.

142 *See* ENGSTROM ET AL., *supra* note 4, at 73.

143 ENGSTROM ET AL., *supra* note 4, at 34, 42 (providing a comparison of the CBP's inability to explain the failure rates of one proprietary COTS iris scanning technology with the SSA's quality assurance tools, made by an attorney-turned-programmer who "developed the flags that [he] wanted to have available as an adjudicator").

leverages or monetizes its relationship with the agency and understanding of the tool in other business relationships. The conflicts of interest underlying capture need not be intentional or explicit. For example, the SEC uses a tool called ARTEMIS to identify suspicious trading through "anomaly detection," working under the theory "that suspicious activity is an outlier" which will not match the pattern of other trading data.[144] If the SEC had procured ARTEMIS from a contractor which had easy access to particular firms' trading data, possibly because of prior business relationships, then those firms' trades could be overrepresented in the data pool and may be less likely to be flagged as suspicious. Meanwhile, other firms' trades could be underrepresented and possibly more likely to be chosen as enforcement targets by the AI. The decision to buy could also facilitate gaming or adversarial learning. For example, imagine that a hypothetical contractor responsible for developing ARTEMIS keeps the internal workings of the algorithm as a trade secret, revealing it to trading firms only for a price. Additionally, we may value giving agencies an implicit "deference incentive" for in-house development whereby agencies would receive softer touch in judicial review of agency action in return for investing in their technical capacity and developing AI tools internally.

On the other hand, we may worry about giving agencies too strong an incentive to develop AI tools in house, especially early in the technological revolution of the administrative state. In the short term, technical capacity is fixed. Where efficiency gains necessitate the development of AI tools, but agencies lack the technical expertise to develop effective AI in-house, careful procurement is likely the best outcome. A deference calculus that automatically applies hard look review to tools developed by contractors may incentivize an agency which is wary of litigation challenges to build those tools in-house despite lacking the capability to do so. This arguably suggests that the additional deference an agency receives when building a tool in-house should be small or nonexistent, to avoid incentivizing the development of bad in-house AIs instead of effective third-party AIs. Another solution may be to grant additional deference based on the agency's diligence in procurement, as signaled by the inclusion of safeguards like algorithmic impact statements, risk management plans, and audit requirements in the procurement contract.

B. AI Explainability

Defined functionally, an explainable AI is one which produces decisions which are also persuasive to human decision makers. Explainability can be

---

[144] *Id.* at 24.

framed in different ways, including "legal" explainability, where a human can parse and use legally valid reasoning from the AI's decision, and "causal" explainability, where a human can deduce (but not necessarily understand) the mechanical functions of the algorithm.[145] Certain AI techniques are more capable of explanation because their structure is based directly on, or aligns with, human cognition, as with decision trees.[146] Other models are much more challenging to explain, such as neural networks, which have complex internal topographies not easily translated into a human mental model.[147] Possible sources of a lack of explainability include complexity, such as where the volume of interdependent inputs, internal calculations, and references to training data exceed ready comprehension.[148] Another source is non-intuitiveness, where an AI's decision rule can be deduced at some level, but human intuition cannot understand why that rule helps to reach a correct decision.[149] Finally, secrecy can also confound AI explanability, particularly in the case of third-party tools or AIs which make decisions potentially subject to gaming.[150]

Explainability is a particular challenge for AI users and developers. Many AIs fail to communicate how they function at any level, and others fall short of abstracting the factors influencing their decisions to a level compatible with human psychology and understanding.[151] For example, machine learning tools often entirely lack causal chains in reaching their decision, making it nearly impossible to map their decisions onto human reasoning.[152] Moreover, there is a dearth of formal theory on measuring the quality of an explanation, in part because there is no consensus on a rigorous definition of what it means to explain or understand in the first place.[153]

Favoring AI explainability may come at the expense of accuracy. Often the accuracy gains that AI promises are at least partially predicated on AI's ability to recognize patterns which humans both cannot find and struggle to comprehend.[154] This tradeoff may not always be worth making, especially given

---

145 Henrik Palmer Olsen et al., *What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration* 8–9, (iCourts Working Paper Series No. 162, 2019).

146 Lisa Käde & Stephanie von Maltzan, *Towards A Demystification of the Black Box — Explainable AI and Legal Ramifications*, 23 J. INTERNET L. 4, 4 (2019).

147 *Id.*

148 COGLIANESE, *supra* note 7, at 45–46.

149 *Id.* at 46.

150 *Id.*

151 *See* Walter A. Mostowy, *Explaining Opaque AI Decisions, Legally*, 35 BERKELEY TECH. L.J. 1291, 1306–07 (2020) (discussing how explainable AI commonly abstracts its decision-making channels only to the point of the weights of low-level technical components, but not high-level human-readable concepts).

152 COGLIANESE, *supra* note 7, at 46.

153 Mostowy, *supra* note 152, at 1307.

154 ENGSTROM ET AL., *supra* note 4, at 28; Käde & von Maltzan, *supra* note 147, at 3–4;

the legal realist contention that human decision making may also be a black box, albeit one with an added layer of rationalization.[155] At the same time, some scholars consider explainability an uncompromisable value for government use of AI, and for government administrative action in general.[156] Explainability can legitimize the agency and its actions.[157] It can also improve the quality of decision making, since a wrongful agency action which is demonstrably irrational is harder to defend than an inexplicable one.[158] If AI undermines the agency's ability to justify its actions, critics argue, it may not be worth adopting at all.[159]

A *Baltimore Gas* analysis may not directly add value to the question of how much deference is due based on AI explainability. The explainability of the AI underlying an action arguably does not implicate the substantive expertise that the agency brought to bear on the problem, since an agency can equally apply its substantive expertise to an explainable AI technique and on an unexplainable one. The policy considerations of the Bazelon-Leventhal debate facially prefer softer look review as AI grows more explainable. Agencies are likely to make better choices when their actions are subject to review for the quality of those decisions, but this review cannot happen unless the agency actually provides reasons for the public to challenge and the court to scrutinize. This transparency is overall legitimizing for the agency and its actions. So, it may be advantageous to allow courts to reach deeper into the substance of agency action when that action is based on a less explainable AI to offset the fear that the AI's inexplicable decisions will be otherwise unchecked. Under this theory, causal explainability of only the mechanisms in the AI is not worth additional deference since it does not fulfill the requirements of agency reason-giving. Instead, legal explainability should be the tethering point for deference.

But this argument starts to break down when we consider the tradeoff between explainability and accuracy in the context of the Bazelon-Leventhal debate. If more complex AIs are less explainable, then granting soft look review to agency action based on explainable AIs will implicitly apply hard look review to agency action based on complex AIs (which are, again, likely to be less explainable). Such a paradigm runs counter to the intuition that courts should defer to substantive agency expertise. To the contrary, the more complicated and difficult

---

Mostowy, *supra* note 152, at 1300–01.

155 COGLIANESE, *supra* note 7, at 46–47.

156 Käde & von Maltzan, *supra* note 147, at 3–4 (noting that the explainability requirement emerged from the need for the public to challenge administrative decisions); *see* Olsen et al., *supra* note 146, at 7.

157 Olsen et al., *supra* note 146, at 7.

158 *Id.* at 7.

159 *Id.* at 4.

to understand an AI is, the more a judge, who is lacking in AI expertise, would be obligated to wade in and assess the agency's action based on the AI's merits.[160] This deference incentive also implies that an agency which is averse to litigation may choose an ineffective, explainable AI to avoid more frequent or expensive litigation. Also, if courts always take a harder look at actions based on unexplainable AI, they effectively ask the agency to explain the unexplainable, imposing a great burden on the agency. We may worry that withholding deference from agencies for using unexplainable AI tools will effectively proscribe the use of more complex, advanced forms of AI, relegating agencies to less effective alternatives.[161] Accordingly, we might argue that AI explainability, as a proxy for AI sophistication, should not impact the deference decision one way or the other. But such a system would have an incentive problem as well, since agencies facing an even-keeled deference regime might want to make needlessly complicated AIs to evade the reason-giving requirement and insulate decisions from challenges. Perhaps a better solution is job-like: give deference with one hand for accuracy (the benefit of sophistication) measured in terms of false positives, false negatives, and other appropriate metrics,[162] and take deference away with the other hand for lacking explainability (the harm of sophistication).[163] Such a standard also aligns better with my interpretation of *State Farm*.[164]

---

160 On the other hand, we may want to draw a distinction between agency expertise in core agency functions, and agency expertise in peripheral functions, such as making AI. So, while an agency's expertise might be implicated in performing cost-benefit of a regulation that the agency promulgates under a statute it administers, it may be that the expertise to develop a complex AI that does the same thing is a different form of expertise not deserving of the same deference.

161 But again, it may be that this tradeoff is justified because explainability and reason giving are such foundational requirements that agencies should not be allowed to deviate from them, regardless of efficiency and accuracy gains.

162 Reuben Binns & Valeria Gallo, *Accuracy of AI System Outputs and Performance Measures*, INFO. COMM'R'S OFF. (May 2, 2019), https://ico.org.uk/about-the-ico/media-centre/ai-blog-accuracy-of-ai-system-outputs-and-performance-measures/ (discussing various measurements for AI accuracy) (last visited Sept. 11, 2023).

163 Such a solution may raise concerns of bias, not explicitly accounted for so far in my deference framework. An AI which is very accurate but not explainable could receive softer look review despite being biased, and despite the lack of reason giving which would aid in uncovering that bias.

164 *See supra* notes 104–106 and accompanying text ("Where the agency obviously does not bring its substantive expertise to bear in acting (as by making a glaring logical error in its reasoning), even though action in good faith would obviously require the use of such expertise, a court is permitted to include that obvious lack of credibility in its ultimate determination on the deference the action is owed."). *State Farm*'s goal is to promote reason-giving, yes, but it is also to promote deference to expertise. A flat deference standard regardless of AI explainability seems incompatible with the former, while additional deference based on explainability seems incompatible with the latter.

C. Data

Data and machine learning are often treated as synonyms. Many AIs acquire their predictive accuracy from training on large volumes of data, which requires physical and digital infrastructure as well as network access and proper cybersecurity.[165] Humans cannot compare to algorithms in terms of utilization of data. We have limited memory capacity and lifespan, and struggle to transfer information effectively.[166] But humans do not usually utilize data in decision making to the same extent that AI does.[167] So, the relevant baseline comparison for algorithms may be data-rich AI decisions versus data-deficient human decisions, or it may be effectiveness in use of data by algorithms versus effectiveness in its use by humans.

Potential problems with relying on data include insufficient data, inaccurate data, poor feature selection, and unrepresentative sample selection. Although there is no bar to creating AI with small data sets, in practice thousands or even millions of observations are required to capitalize on the promises of machine learning, so agencies must be sure to collect sufficient data.[168] Inaccuracies due, for example, to measurement errors or incorrect reporting, can also decrease the effectiveness of AI in deducing patterns from the data set.[169] Feature selection refers to the selection of input variables to an algorithm.[170] Failing to include features in a dataset which are relevant to the prediction the AI is asked to make can undermine model accuracy.[171] Conversely, including features in a dataset which are not useful for an algorithmic prediction can increase the risk that the AI will fit too closely to the training data, requiring exponentially more observations in the data set to reach the same accuracy.[172] Also, data based on a poorly selected sample will fail to generalize because the algorithm extracts

---

165 COGLIANESE, *supra* note 7, at 41.

166 *Id.* at 42.

167 I would argue that extrapolating from data complicates AI and makes it less predictable and deterministic. Where sufficient accuracy can be achieved otherwise, as by simulating physical phenomena according to known laws, adding data is simply a recipe for occasional incorrect, inexplainable results. Take the PHASE software at the FDA discussed by Professor Mason Marks. Mason Marks, *Automating FDA Regulation*, 71 DUKE L.J. 1207, 1227–29 (2022). Developers and scientists can use deterministic laws of nature to answer the question of whether particular molecules will bind to a given receptor, and I think it would be a mistake to use AI, which can introduce unpredictable deviations from those natural laws, where the simplicity, maintainability, and accuracy of a simulation is sufficient.

168 David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 678–79 (2017).

169 *Id.* at 679.

170 *Id.* at 700–01.

171 *Id.*

172 *Id.*

patterns from a group of individuals that looks different from the population to which it is eventually applied.[173] Even when a dataset meets all these characteristics, storing the data can still pose cybersecurity, privacy, and data maintenance concerns. Cybersecurity is a particular concern in governmental use of data, since the information guarded can often be both more valuable (the government handles large aggregations of personally identifiable information ("PII")) and less secure (the government has limited resources and cybersecurity expertise) than private sector databases.[174] This concern intersects with privacy, as personal data can be sold, stolen, or even inferred from incomplete data with high accuracy.[175] Finally, storing and maintaining data long-term is expensive, as physical representations of data decay quickly, and digital tools to retrieve the data become obsolete.[176]

Under *Baltimore Gas*, the presence and quality of data in training an AI implicate an agency's substantive expertise underlying an agency action based on that AI. First, much like the procurement decision for the AI, the way the agency acquires the data for the AI would affect deference. If the agency acquires the data itself in the course of performing core agency functions, then the agency has exerted expertise in feature selection and ensuring the sufficiency, accuracy, and representativeness of the data.[177] At the opposite extreme, if an agency purchased the entire AI system, it would be hard to argue the data implicates any of the agency's expertise. Slightly better in terms of expertise are cases where the agency buys the data set, since the agency may then at least inspect the data for sufficiency. Further along the expertise spectrum are situations where an agency procures the data set itself, but the data does not directly reflect the agency's expertise. For example, the USPS's autonomous driving AI requires data collected from sensors on driving cars.[178] The USPS can collect this data itself, but since the agency's expertise ostensibly lies in delivering packages, not the act of driving, then that expertise is not as effective a barometer of accuracy, feature selection, and representativeness of the data set.[179] Beyond acquisition, however, a judicial inquiry into the data itself is likely

---

173 *Id.* at 713.

174 COGLIANESE, *supra* note 7, at 43.

175 *Id.* at 44–45.

176 *Id.* at 41–42 (noting that most AI techniques, once trained upon data, no longer require access to that data, unless the model is to be retrained).

177 ENGSTROM ET AL., *supra* note 4, at 23 (describing how the SEC's ARTEMIS tool is built on a dataset of 8-K forms filed with the SEC).

178 *Id.* at 66.

179 *See* U.S. Postal Service, *Postal Facts: Innovation in the Mail*, https://facts.usps.com/innovation/#fact701 (offering empirical evidence suggesting that UPSPS is ideally situated, at least for a public sector actor, to develop an autonomous vehicle) (last visited Nov. 26, 2023). Perhaps USPS is ideally situated, at least for a public sector actor, to develop an autonomous vehicle.

undesirable. If judges must assess deference based on not just how the agency acquired data, but whether that data is of sufficient quality for use in AI, then this threshold determination is already so substantive that it is as if the court is performing hard look review. Such a system dissipates any judicial efficiencies from deferring and provides the court an opportunity to inject discretion where the correct outcome is deference.

The Bazelon-Leventhal debate factors also suggest adjusting deference to the agency's use of data. When an agency collects data itself, it is arguably fulfilling an executive role to a greater degree than it would be by merely purchasing data, emphasizing the separation of powers concern motivating soft look review. When the agency collects the data through functions implicating its substantive expertise, the court's relative inexperience also counsels a softer look. Indeed, data which the agency collects within its area of expertise may be the most socially desirable data set for the agency to use because the agency is accountable for that data and its quality, leading to better AI and better agency decisions. Accordingly, it may be beneficial to offer a deference incentive for agencies to use such data as it is. Otherwise, the agency would be required to satisfy additional hurdles in performing its core functions to collect data worthy of deference, which would impose a heavy burden on the agency in performing those core functions and would incentivize acquisition. Data privacy, security, and erosion also influence the deference decision through the explainability concern discussed in Section B, above. Where data is more private or more valuable, the importance of transparency increases, so the court should take a harder look to ensure that the agency is adequately stewarding and protecting it.

D. Human in the Loop

Another challenge in agency implementation of AI is where and how humans should be involved in the algorithmic decision-making process (the "human in the loop" problem).[180] Humans can perform many different roles as part of systems containing AI. Humans can serve "corrective" roles to improve the system's performance in three ways: by correcting factual errors in an AI's decision, by tailoring an AI's otherwise correct decision to a particular context where it is erroneous, and by correcting for systemic bias which has led the algorithm to a statistically accurate but socially unacceptable result.[181] Humans in the loop can also perform "resilience" roles by functioning as a backstop to minimize the harm from bad outcomes, such as when a human wrests control

---

[180] *See* Rebecca Crootof et al., *Humans in the Loop*, 76 VAND. L. REV. 429, 438–39 (2023).

[181] *Id.* at 473–74.

from an autonomous vehicle to avoid an accident.[182] They may be included in the loop to fulfill "justificatory" roles by justifying decisions, particularly when the AI's decision-making process is difficult to explain.[183] Humans can also serve "dignitary" roles, affording those affected by action based on algorithms respect and individualization.[184] Humans in "accountability" roles serve to assuage the concern that agencies will delegate difficult decisions to algorithms so as to avoid taking responsibility.[185] Humans can also serve "stand-in" roles, filling in for regulators in the abstract sense,[186] and "friction" roles, slowing the overall operation or adoption of algorithms in a beneficial way.[187] Humans in the loop may occupy "interface" roles, helping users interact with the algorithm.[188] Finally, humans may simply fulfill a "warm body" role to avoid technology displacing them from their employment.[189]

The law can place varying burdens on agencies to add humans to AI loops. In some use cases, the law mandates that a human be included in the loop, as with a new Colorado law requiring that a human be placed in the loop wherever the government uses facial recognition with a legal effect.[190] In others, the law merely incentivizes adding a human element, such as by allowing procedural challenges to decisions made without a human in the loop, which would otherwise be blocked.[191] An additional possibility, not currently in use, would be for the law to disincentivize or prohibit adding a human to the loop.[192] Such a constraint might exist as part of an effort to prevent the harmful effects of discretionary bias.[193]

---

182 *Id.* at 478.

183 *Id.* at 478–79.

184 *Id.* at 480–81; Meg Leta Jones, *The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood*, 47 SOC. STUD. SCI. 216, 230 (2017) (describing a human in the loop as a fundamental right recognized in much of Europe).

185 Crootof et al., *supra* note 181, at 482–83. The authors observe that a more cynical reading of the accountability role is to provide a "liability sponge" to take the fall for bad results, such as the tendency of Tesla's autopilot to hand off control to a driver before a crash so that the autopilot is not running at the time of (but no less responsible for) an accident. *Id.* In the context of agency action, this could look like an agency adopting an AI which produces a correct result in 99% of cases, and a catastrophic result in that remaining 1%. Regardless of whether adopting the AI leads overall to efficiency gains, the AI will almost certainly produce gains if the agency is allowed to offload accountability to a single human in the loop in an accountability role, who can then be cheaply fired and replaced.

186 *Id.* at 484.

187 *Id.* at 484–85.

188 *Id.* at 487.

189 *Id.* at 485–86.

190 *Id.* at 448; COLO. REV. STAT. § 24-18-303 (2022).

191 Crootof et al., *supra* note181, 449–50.

192 *Id.* at 458–59.

193 *Id.* at 459 (stating that this rationale was a basis for many mandatory sentencing laws).

The dangers of AI without human intervention include failing to fulfill the functions that humans can bring to the loop. Without humans in corrective roles, AI errors may increase. Without humans in resilience roles, AI failures may lead to greater harms. Without humans, we may not have sufficient insight into AI decisions to understand why decisions are made (justificatory roles). We run the risk of infringing human dignity (dignitary roles), failing to allocate responsibility for system failures (accountability roles), proliferating AIs too quickly (friction roles), and adding unnecessary challenges to interfacing with AI systems (interface roles).

On the other hand, the humans we place in the loop are not infallible, and overreliance upon them could undermine the effectiveness of an AI system. The short-term working memory of humans can handle about four variables at once, forcing us to block out (potentially relevant) information to make decisions that an AI can use.[194] Humans, unlike algorithms, suffer from fatigue and aging, which reduce alertness and situational awareness and increases reaction times, memory lapses, and the risk of errors and accidents.[195] Humans can also be impulsive, taking premature action without foresight, and perceptually inaccurate, leading to poor decisions.[196] They are also prone to a number of biases, including racial and gender biases, the endowment effect (valuing things in one's possession more than those who do not have them would value them), loss aversion (fear of losing something one has is greater than the desire to acquire the same thing if one is not already in possession of it), system neglect (misgauging the importance of signals compared to their impact on the system that generated them), hindsight bias ("Monday morning quarterbacking"), availability bias (the assumption that examples which come to mind easily are prevalent than those that do not), confirmation bias (the tendency to search for and favor information confirming existing beliefs), framing bias (altering perceptions of risk based on the relative prominence of information about gains or losses respectively), anchoring bias (making estimates of the unknown based on modifications to an initial anchor), and susceptibility to over-persuasion (as through gruesome language).[197]

Does placing a human in the loop implicate an agency's substantive expertise per *Baltimore Gas*, warranting judicial deference? It depends on the human and their role in the loop. Coglianese's concerns with human decision makers show that an additional human is not always beneficial.[198] A human brings more

---

194 COGLIANESE, *supra* note 7, at 9–10.
195 *Id.* at 11–13.
196 *Id.* at 13.
197 *Id.* at 15–21.
198 *Id.* at 21–22.

expertise to bear on an action when the human is experienced and qualified in the specific decision the AI is making. For example, the SSA's Insight program to provide quality assurance on a written decision includes, as a human in the loop, an ALJ, who is an expert at making that determination.[199] But even if the human brings expertise to bear on the same problem the AI is solving, the agency's expertise may not be implicated if a loop performs an action which is not part of the agency's expertise. For example, the USPS's autonomous driving AI contains a driver as a human in the loop.[200] Though the human is an expert at the problem the AI is trying to solve (i.e., driving), the agency's expertise at delivering mail is not implicated per se in expertise in driving. The human's role also influences the extent to which the human's presence implicates the agency's substantive expertise. A human in a corrective role likely brings the agency's expertise to bear because the human's expertise is factored into every decision the AI makes. Humans in resilience, justificatory, or dignitary roles must leverage their expertise where something goes wrong with the AI decision or where an action is challenged. Humans in accountability, friction, warm body, and interface roles do not bring expertise to bear on the eventual decision that the agency makes, so their presence in the loop may not be a cause for softer look review.

The Bazelon-Leventhal factors also tend to favor additional deference to agency action based on AI with a human in the loop. Where the added human in the loop brings expertise, courts unversed in the subject of the agency's action should be more inclined to defer. Humans in justificatory roles can promote transparency and facilitate reason-giving, lessening the need for hard look review. But a note of caution: if the level of deference given to the agency action does not depend on the loop and the human's expertise and role, then agencies may be incentivized to add ineffective humans to loops to gain additional deference, without actually achieving the additional supervision that supposedly warrants the deference in the first place. Various flaws of human and AI "hybrid" systems, such as Coglianese's list of human shortcomings, automation bias, and "skill fade" (deteriorating human abilities where they are not used), could undermine the system as a whole when compared to a machine-based system alone.[201] So, while an expert human in an effective role in an appropriate loop should indeed receive additional deference from a court, this inquiry is difficult and fraught with the possibility of error. We might question whether assessing the appropriateness of the human to the loop is even worth the candle, and instead simply refuse to adjust deference based on human involvement

---

199 ENGSTROM ET AL., *supra* note 4, at 40.

200 *Id.* at 66–67.

201 *See* Crootof et al., *supra* note 181, at 437, 468–69 (discussing the "MABA-MABA" (men-are-better-at, machines-are-better-at) trap in the context of hybrid systems).

unless the agency demonstrates the value of the human in the loop by some heightened standard like clear and convincing evidence.

E. Use Case

Should the deference inquiry care what the agency uses the AI to do? We can break out use cases in at least two ways. First, we can divide use cases by the stakes of the action which will be based on the algorithm.[202] For example, an AI which performs a purely internal agency staff function is a low-stake use case (though such a use case would typically not find itself under judicial review). At the opposite extreme, a high-stake use case impacts individual liberty or has high financial stakes for regulatory targets. Second, we can categorize use cases based on the kinds of rights they implicate: adjudicatory use cases, enforcement use cases, and "other" use cases. Other use cases include use for regulatory analysis, such as the FDA's adverse drug event detection AI,[203] and for public engagement, such as the CFPB's consumer complaint processing AI.[204] The question of stakes correlates with the rights-based categorization, since adjudicatory AIs are likely higher risk than enforcement AIs, which can be higher risk than other use cases. Still, having both divisions is useful because use cases in the same category are likely to have similar effects. AIs in adjudicatory use cases can affect decision quality for better or worse, and, if we accept the right to a human decision maker, can undermine hearing rights.[205] Conversely, if regulatory bodies or law enforcement agencies use AI tools to choose enforcement targets, they could potentially select incorrect targets and subject those targets to undeserved litigation costs.[206]

---

202 COGLIANESE, *supra* note 7, at 72–74.

203 ENGSTROM ET AL., *supra* note 4, at 55.

204 *Id.* at 61–62.

205 *Id.* at 83–84.

206 Should we worry about the risk of poor prioritization of targets for enforcement? From the perspective of the party against whom enforcement is sought, it may seem inequitable to be subject to an enforcement proceeding even though a more egregious offender continues unaccosted. But since the target is not above enforcement, this concern is not particularly persuasive. From the agency's perspective, poor choice of targets may be inefficient since agencies can only bring limited enforcement actions with their finite resources. But *Heckler* warns that we should be wary of courts interjecting their opinion on agency enforcement decisions and prioritizations rather than the instant case. Heckler v. Chaney, 470 U.S. 821, 831 (1985) ("[A]n 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 . . . . This recognition of the existence of discretion is attributable in no small part to the general unsuitability for judicial review of agency decisions to refuse enforcement."). It could be that the agency rationally chooses to spend its finite resources bringing more enforcement actions, rather than spending those resources on excessive

The type of use case, though not the stakes, affects the substantive expertise the agency brings to bear on an action per *Baltimore Gas*. While we may like to assume that agencies leverage more expertise on higher stakes decisions, it would be naïve to assume that the stakes of the action are an effective proxy for an agency's actual use of substantive expertise. On the other hand, stakes may be correlative with other protections, such as the additional procedural protections afforded by the APA in formal adjudications compared to informal ones.[207] One could argue that where the agency has satisfied many procedural checks, a court should grant the agency a break through additional deference. This argument is unconvincing, because post-*Vermont Yankee*, procedural and substantive review are complements, not substitutes.[208] If stakes rise and procedural requirements increase correspondingly, that is no reason to reduce substantive review. So, even where stakes correlate with procedure, that is not a cause for a grant of deference. The category of the use case may also implicitly demonstrate the extent to which the agency uses its substantive expertise if the category aligns with the agency's functionality. For example, a law enforcement agency like CBP may inherently leverage expertise in using an AI to perform a law enforcement function that it would not leverage in a use case from the "other" category such as public engagement. Such categorical alignment between agency and use case should merit additional deference.

This discussion of the use case in which the AI is deployed has centered entirely on the use case, and not at all on the AI. I argue that the use case implicates a fixed level of agency expertise regardless of how the agency decides to accomplish that use case. If using AI to perform the use case influences the level of expertise the agency brings to bear in fulfilling the use case, then it is likely because the choice to use AI abdicates some of the expertise which would otherwise be utilized by the agency if it were to procure AI tools instead of building them in-house. Characteristics of the AI which affect the expertise brought to bear are, in my opinion, better addressed directly through other factors in this framework, such as in the deference determination based on the "buy or build" decision, rather than being shoehorned into the use case assessment. That is not to say that the use case is unimportant to the deference

---

prioritization of targets.

[207] *See* 5 U.S.C. §§ 554(a), (c)(2) (applicable to formal adjudications which are "on the record after opportunity for an agency hearing" and providing for oral presentation and cross examination).

[208] Vermont Yankee Nuclear Power Corp. v. Nat. Res. Def. Council, Inc., 435 U.S. 519, 545 (1978) ("Respondent NRDC argues that § 4 of the Administrative Procedure Act, 5 U.S.C. § 553 (1976 ed.), merely establishes lower procedural bounds . . . . [O]ur decisions reject this view."); *id.* at 549–50 ("We now turn to the Court of Appeals' holding 'that rejection of energy conservation on the basis of the "threshold test" was capricious and arbitrary'. . . .").

decision under *Baltimore Gas*. Instead, the use case is to be assessed independently of the agency's means for achieving that use case. Thus, although the use case affects the level of expertise the agency brings to bear, the use of AI in accomplishing that use case does not.

The use of AI within a use case, however, does directly affect the Bazelon-Leventhal factors of deference in substantive review. First, as with the *Baltimore Gas* analysis above, the use case is important in isolation from the method of achieving that use case. As stakes grow, so do the risks associated with potential agency capture, the need for supplemental checks on agency power, and the overall value of transparency, all of which point toward hard look review. Similarly, the value of predictability increases, and the value of agency flexibility correspondingly decreases, undermining an advantage of soft look review. The categorical assessment of use cases is similar. For adjudicatory use cases, the concern with separation of powers matters less because the agency is acting as a decision maker, so the value of a better and more transparent decision through judicial review is at its highest. For enforcement use cases, courts should be more deferential to the agency's discretion not to bring enforcement actions. This squares with the policy concern that the burden on the agency to justify not just the enforcement action but the decision to bring *this* action out of all possible actions is too great, resulting in debilitating ossification at the agency, and eroding the separation of powers. For an example of weighing both stakes and category together, consider the SEC's ARTEMIS tool.[209] Its use case is choosing targets for regulatory enforcement by identifying suspicious trades. The enforcement category of use cases should get softer look review because courts defer to agency decisions of who to prosecute, but the repercussion is large financial penalties, so the relatively high stakes increase the appropriate hardness of review slightly.

Unlike the *Baltimore Gas* inquiry, the decision to use AI to achieve a use case does implicate policy concerns with judicial deference. The hardness of review should be tailored to the overall suitability of AI for the task by considering whether the potential perils of AI are more severe in that particular use case. These perils include concerns with delegation and accountability, procedural due process, and reason-giving, privacy, and bias.[210] For example, if an agency use case would have outsized effects if that agency's action were riddled with algorithmic bias, then the decision to use AI to perform the use case should be subject to harder look review because the value of transparency is higher in this setting. Note that this additional scrutiny is appropriate regardless of the particulars of the AI and reflects the policy desire to use hard look review to

---

209 ENGSTROM ET AL., *supra* note 4, at 23–24.
210 COGLIANESE, *supra* note 7, at 50.

force agencies to make better decisions and supplement existing checks on agency action. That is, deference should be adjusted both for an unexplainable AI and for a use case where a potentially unexplainable AI would cause bigger concerns. Both of these deference adjustments can be made independently by inspecting the AI and the use case respectively. Where the perils of governmental use of AI would have an outsized impact on a particular use case, review of actions based on AI should be less deferential.

F. Model Implementation

As explored above, AI techniques are ever-increasing and vary widely in sophistication and effectiveness. I do not contend that as part of deciding whether to use hard look review, a court should involve itself in, for example, an agency's decision to use an AI based on generative adversarial networks instead of one based on Markov chains. Such a determination asks too much expertise of the court for too little a reward. But certain technical decisions in choosing and implementing an AI implicate procedural consequences that courts cannot neglect. Any effort to determine the deference due to every AI technique in this paper would be both long-winded and out-of-date as soon as written, but a good governing principle is that courts should rarely tailor deference to the specifics of an AI unless the agency's choice of model thwarts the agency, the public, and past courts in understanding the AI as applied.[211] To illustrate, we will consider two example characteristics of AI models: the supervised/unsupervised learning distinction, which does not influence the deference decision, and the "online/offline" distinction, which plays an important role in the choice between hard- or soft look review.

1. Supervised and Unsupervised Learning

In the training phase of supervised learning, an AI is provided data that includes as a label the expected output corresponding to each sample input.[212] The AI then adjusts its parameters to correctly predict labels for unseen inputs.[213] In the training phase of unsupervised learning, an algorithm must adjust its parameters without the assistance of labeled ground truth answers.[214] Unsupervised learning is a useful approach when experts are unsure of common

---

[211] This concern dovetails with explainability, discussed *supra*, but the two do not always go hand in hand. Again, to be clear, the only question this and the other factors weighs on is whether to perform harder or softer look review. If a court chooses to perform hard look review, then it absolutely should inspect the specifics of the AI.

[212] COGLIANESE, *supra* note 7, at 45–46.

[213] *Id.*

[214] *Id.* at 25.

properties within a data set.

All else equal, *Baltimore Gas* is indifferent to whether an AI is supervised or unsupervised. Any argument to the contrary would be an assertion by the court either that one form or the other always implicates more of the agency's expertise, or that the court is capable of determining whether a supervised or unsupervised model requires more of the agency's expertise in a given set of circumstances. The first position, that either supervised or unsupervised learning always requires more expertise, is clearly false: there are complex and technical AI techniques using both supervised and unsupervised learning suitable to different problems, just as there are simple, straightforward algorithms using each.[215] The second position, that a court can determine as part of a threshold deference inquiry whether supervised or unsupervised learning requires more expertise, would undermine judicial efficiency and the other benefits of deference. Instead, while a court performing a hard-look review of an agency action based on AI should consider the complexities of the model being made, the court should base its determination to perform that hard-look review on more digestible considerations.

The Bazelon-Leventhal factors point to the same conclusion. Details of an AI model like the decision to use a supervised or unsupervised technique do not impact the separation of powers or need for additional checks on agency action. Judges do not have more expertise in assessing one AI model over another, nor is the burden on the agency of judicial review greater when the court reviews one model over another. The desire to motivate an agency to make good choices does not vary based on the specifics of the AI model used to make those choices, nor does the fear of agency capture. One may argue that AI explainability is negatively correlated with the choice to use unsupervised learning, and therefore unsupervised learning should be subject to harder-look review. But since explainability is considered independently of the model, it would be a mistake to reconsider explainability in the more technical context of the specific implementation of the AI, where a court is more likely to make an error.

2. Online and Offline Learning

"Offline" algorithmic learning refers to training a model on a data set before

---

[215] *See* Julianna Delua, *Supervised vs. Unsupervised Learning: What's the Difference?*, IBM CLOUD BLOG (Mar. 12, 2021), https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning (explaining how clustering is a straightforward unsupervised learning technique, while dimensionality reduction is a challenging one, and regression is a simple supervised learning technique, while neural networks are an exceptionally complex one).

deployment.[216] That initial data set is all the information the deployed model will ever learn from.[217] If the model grows ineffective, or if the developer acquires some data which it believes would be useful in the model, the model must generally be retrained, using a new or augmented data set.[218] In contrast, "online" learning allows a model to continuously update with new data, without ever retraining on the old data.[219] Advantages of online learning include the ability to adapt to new data without costly and time-consuming retraining, along with the ability to forego the expenses of storing the old data.[220] On the other hand, it may be difficult to keep current with an online algorithm's performance, since accuracy, tendency toward bias, and other metrics can change as the model consumes new data.[221]

All else equal, an offline algorithm deserves more deference than an online algorithm. Under *Baltimore Gas*, an agency has presumably brought expertise to bear in building and/or assessing an AI as of time of completion. With an offline-learning model, that expertise remains reflected in the AI as long as it is in use. With an online-learning model, the AI may immediately begin drifting away from the instantiation that survived the agency's expert review, obviating that expertise. The Bazelon-Leventhal concerns again point to the same conclusion, since the value of transparency skyrockets when the AI that an agency is using as a basis for its actions can evolve away from the version that stakeholders reviewed (the agency, through deploying the AI; the public, through notice and comment; and the court, through prior challenges to agency action based on that AI). For example, when Microsoft debuted a bot account on Twitter which used online learning to improve at generating natural human speech through engagement with users, the account quickly began mimicking racist comments by other users in a development that could not have been foreseen (from a technical perspective, at least) at the time Microsoft released the bot.[222]

Online algorithms also increase the concerns with agency capture through gaming by permitting regulatory targets to feed the AI new data to corrupt not just its outputs but its entire decision-making process. For example, consider the example from the ACUS report of patent applicants manipulating the images in

---

216 *See generally* Max Pagels, *What Is Online Machine Learning*, MEDIUM (Apr. 20, 2018), https://medium.com/value-stream-design/online-machine-learning-515556ff72c5.

217 *Id*.

218 *Id*.

219 *Id*.

220 *Id*.

221 *Id*.

222 Daniel Victor, *Microsoft Created a Twitter Bot to Learn from Users. It Quickly Became a Racist Jerk*, N.Y. TIMES (Mar. 24, 2016), https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html.

their applications with random noise, which, while unnoticeable to the human eye, can trick machine learning models into misclassifying images and funneling applications to examiners in subfields with higher grant rates.[223] Applicants could similarly and more predictably manipulate an online-learning classifier by, for example, first filing applications with a digital "watermark" such as a particular kind of image noise, which would be correctly classified into a certain subfield. Then, once the algorithm has "learned" the classification pattern, the applicant could use that watermark in future applications to gain similar treatment. This increased risk of gaming heightens the Leventhal concern.

Both the *Baltimore Gas* and Bazelon-Leventhal inquiries suggest harder-look review for online algorithms than offline and no change in deference between supervised and unsupervised learning. This aligns with my position that, in general, a reviewing court would do more harm than good by attempting to tailor deference to a model's particulars, because such review is too detailed and difficult to be a threshold inquiry. The exception, where difficult review becomes worthwhile, is where the model's implementation undermines the ability of the public to engage with the AI, by, for example, changing the AI's parameters after the public has had its opportunity to pass comment. With such AIs, courts should engage in harder-look review.

CONCLUSION

As part of the trend toward algorithmic governance, agencies use many different AI techniques to solve many different problems with many different degrees of success. To respond to litigation stemming from the rising tide of agency AI use, courts require a framework for determining the level of judicial deference due to agency action based on AI. Despite the hallowed status of *State Farm*'s hard-look review of agency action, the throughline of the Court's jurisprudence on arbitrariness review is deference. *Baltimore Gas* most clearly articulates the Court's position that the level of deference a court gives to agency action should be tied to the substantive expertise that the agency utilizes when taking the action. While it is possible to read *State Farm* as consistent with the *Baltimore Gas* standard, it is also important to remember that, to the extent that *State Farm* sanctions a lack of deference to the agency at all, it is an empirical outlier.

Accordingly, this paper's main contribution is identifying a set of factors that influence the level of deference a court owes an agency in reviewing action that is based on AI, and analyzing those factors both under the expertise-based

---

[223] ENGSTROM ET AL., *supra* note 4, at 86.

deference standard of *Baltimore Gas* and the policy considerations of hard-look review raised by Judges Bazelon and Leventhal on the D.C. Circuit in the 1960s. First, agencies should get less deference from courts when they buy AI solutions, and more deference when they build them, because the agency leverages more expertise in building an AI and is more accountable for its performance. Second, AIs which are explainable better fulfill agencies' reason-giving obligations, facilitating better judicial review and warranting more deference. Third, the data, if any, that the agency uses to train an AI influences the deference decision, both through the way the data is acquired and the diligence with which it is kept and used. Data implicates the minimum level of agency expertise when the agency never sees the data, as by procuring a completed AI, where more expertise is used when the agency collects the data itself, and the maximum expertise is used when the agency collects the data while performing core agency roles. Moreover, when the data stored is more valuable or implicates additional privacy concerns, the need for transparency counsels more substantive review to ensure the agency is securing and storing the data effectively. Fourth, when a human is included in the loop, additional deference may be due depending on the role fulfilled by the human in the loop. Expert humans in corrective roles implicate an agency's expertise more than inexperienced humans and those whose position in the loop exists solely to preserve their employment. Fifth, the use case to which the AI is applied can implicate more or less expertise based on whether the agency is acting as an enforcer, an adjudicator, or some other role, and the stakes of the use case can alter the value of transparency. Finally, in most cases judges should avoid wading into the specific implementation of the AI model at the stage of assessing deference in order to avoid concerns with courts' institutional competence at designing artificial intelligence. The exception is that where the model's implementation obviates procedural checks and prior review of the model, courts should scrutinize the model more closely through hard-look review.

The stakes of choosing the appropriate standard of review are the agency's cost of acting, juxtaposed against the public's ability to hold the agency accountable. AI promises a more efficient government, but it also threatens an opaque and potentially biased one. On the one hand, courts must be humble enough to admit when they do not know enough about AI to supplant an agency's determination. On the other, courts must be sharp enough to discern when an agency shirks responsibility for its actions behind an ineffective or obscuring algorithm. The proposed framework offers courts a tool to walk that line appropriately by discerning AIs worthy of deference from those that require careful judicial scrutiny.