# JUDICIAL REVIEW OF COMMISSIONER HAL 9000

Eric Leis*

CITE AS: 7 GEO. L. TECH. REV. 179 (2023)

TABLE OF CONTENTS



* J.D., Notre Dame Law School '22. I thank the Honorable Kenneth Ripple for his helpful comments with this article, John Alsterda at Stanford for introducing me to AI, and the many Notre Dame law professors who indulged my interest in AI and the law including Bruce Huber, Jeff Pojanowski, and Nicole Garnett.

## INTRODUCTION

The federal government is expanding its use of artificial intelligence (AI) to perform key human tasks. But the Administrative Procedure Act (APA), enacted in 1946, is ill-equipped to handle the use of AI by administrative agencies.[1] Given its dynamic nature, approval of AI software in the rulemaking process is unlikely.[2] At the same time, the AI black box problem means that courts are unable to evaluate the "reasoned decisionmaking" of an AI adjudicator in the same way that they currently review determinations by Administrative Law Judges (ALJ).[3] Despite these challenges, the federal government continues to encourage the implementation of AI technologies by administrative agencies. This places the federal judiciary on an AI collision course.

Sooner or later, a federal court will find itself in the difficult position of evaluating an agency decision heavily reliant upon—or made entirely by—AI software.[4] In the absence of Congressional overhaul of the APA, that court will face the arduous task of crafting a new judicial standard appropriate for reviewing AI determinations that still fits within the confines of the APA and a century of case law. This Paper aims to develop a judicial standard for courts to apply when reviewing factual determinations made by agency AI during formal adjudication.

---

[1] The Administrative Conference of the United States (ACUS) defines artificial intelligence as systems that "tend to have characteristics such as the ability to learn to solve complex problems, make predictions, or undertake tasks that heretofore have relied on human decision making or intervention." Admin. Conf. of the U.S., Administrative Conference Statement #20: Agency Use of Artificial Intelligence, 86 Fed. Reg. 6616, 6616 n.1 (Jan. 22, 2021) [hereinafter Statement #20].

[2] *See* David Engstrom & Daniel Ho, *Algorithmic Accountability in the Administrative State*, 37 YALE J. REG. 800, 842 (2020) ("[E]ach model may be distinct so that the model reviewed at one stage (notice and comment) may already be substantively quite different when deployed . . .. These challenges become more acute as agencies adopt more advanced forms of machine learning that are dynamic in nature.").

[3] The black box problem "occurs whenever the reasons why an AI decision-maker has arrived at its decision are not currently understandable . . . because *the system itself* is not understandable." Jordan Joseph Wadden, *Defining the Undefinable: The Black Box Problem in Healthcare Artificial Intelligence*, J. MED. ETHICS, at 764, 767 July 21, 2021.

[4] Scholars have predicted that "an AI application may someday be capable of predicting the likely language, legal authority, and evidentiary basis of a decision," thus "convert[ing] an ALJ's role from drafting to simply signing an automated body of text." Engstrom & Ho, *supra* note 2, at 812–13.

A perfect storm is brewing in the vicinity of judicial review of unfavorable disability determinations by the Social Security Administration (SSA). The SSA has openly embraced AI technology.[5] It uses AI to support human decision-makers at several different stages in the disability claim process.[6] The SSA also frequently finds itself in federal court when claimants petition for judicial review of unfavorable disability determinations.[7]

When federal courts review SSA disability determinations, they apply the "substantial evidence" standard which dictates that agency decisions be upheld as long as the ALJ demonstrates "reasoned decisionmaking" and the conclusion is based on "such relevant evidence as a reasonable mind might accept as adequate to support a conclusion."[8] The "substantial evidence" standard, therefore, will provide the basis for a new AI standard, called the "substantial evidence for AI" standard, to be developed in the following pages. This new standard provides the framework for judicial review of agency factual determinations in formal adjudications, in general, and social security disability determinations, specifically.

Part I of this Paper explains how different branches of the federal government have encouraged agencies to adopt AI and how the agencies have responded. Part II discusses the process for obtaining disability from the SSA and the role of AI in that process. Part III presents the "substantial evidence" standard as it functions today and explains why the SSA's AI decisions are ill-suited for the standard. Part IV revises the "substantial evidence" standard to create a "substantial evidence for AI" standard which can be followed when reviewing agency AI decisions.

## I. THE IMPLEMENTATION OF AL BY THE FEDERAL GOVERNMENT

AI is no longer the exclusive property of Silicon Valley. In recent years, Washington D.C. has taken significant strides towards implementing AI across the federal government. The extent to which elected leaders and appointed officials have gone to creating an AI-driven federal government portends a future justiciable dispute between a human and agency AI. Part I explains the efforts made by Congress and the White House to encourage the use of AI in the government and the applications put into place by administrative agencies.

---

[5] *See* discussion *infra* Section II.
[6] *See id*.
[7] *See* discussion *infra* Section II. A.
[8] Biestek v. Berryhill, 139 S. Ct. 1148, 1154 (2019) (quoting Consol. Edison Co. v. NLRB, 305 U.S. 197, 229 (1938)); Allentown Mack Sales & Serv. v. NLRB, 522 U.S. 359, 374 (1998).

### A. Encouragement of AI by Congress

Congress has taken a noticeable interest in AI in recent sessions. In any given year between 2011 and 2016, members of Congress mentioned AI no more than nine times.[9] That all changed in 2020, when reported discussions of AI skyrocketed to 129 in that year alone.[10] The 116th Congress, which convened from January 2019 to January 2021, enacted at least four laws "focus[ing] on AI or includ[ing] AI-focused provisions."[11] Some of the legislation from the 116th and 117th Congresses focused on AI from a national security perspective.[12] Other proposed bills addressed facial recognition technologies and the use of AI in industries like energy, healthcare, and law enforcement.[13]

One piece of legislation, the AI in Government Act of 2020, codified the AI Center of Excellence within the General Services Administration. The AI Center of Excellence is tasked with "facilitat[ing] the adoption of artificial intelligence technologies in the federal government" and "improv[ing] cohesion and competency in the adoption and use of artificial intelligence within the Federal Government."[14] In performing this role, the AI Center of Excellence advises agency leaders on the adoption of AI and develops partnerships between the government and private industry to ensure the former has access to the latest innovations.[15]

### B. Encouragement of AI by the President

---

[9] Laurie A. Harris, CONG. RSCH. SERV., R46795, ARTIFICIAL INTELLIGENCE: BACKGROUND, SELECTED ISSUES, AND POLICY CONSIDERATIONS 22 (2021).
[10] *Id.*
[11] *Id.* at 23.
[12] *See id.* at 24–25 (discussing the National Defense Authorization Act (NDAA) and Identifying Outputs of Generative Adversarial Networks Act); *see also Summary of AI Provisions from the National Defense Authorization Act 2021*, STANFORD UNIVERSITY HUMAN-CENTERED ARTIFICIAL INTELLIGENCE, https://hai.stanford.edu/policy/policy-resources/summary-ai-provisions-national-defense-authorization-act-2021[https://perma.cc/XT8T-JMAL] (last visited Apr. 10, 2022); *Summary of AI Provisions from the National Defense Authorization Act 2022*, STANFORD UNIVERSITY HUMAN-CENTERED ARTIFICIAL INTELLIGENCE, https://hai.stanford.edu/summary-ai-provisions-national-defense-authorization-act-2022 [https://perma.cc/CSC7-GYLL] (last visited Apr. 10, 2022).
[13] *See* ARTIFICIAL INTELLIGENCE BACKGROUND, *supra* note 9, at 25.
[14] *Id.*
[15] Kathleen Walch, *How the Federal Government's AI Center of Excellence is Impacting Government-Wide Adoption of AI*, FORBES (Aug. 8, 2020, 01:00 AM), https://www.forbes.com/sites/cognitiveworld/2020/08/08/how-the-federal-governments-ai-center-of-excellence-is-impacting-government-wide-adoption-of-ai/?sh=79e94db6660e#14912256660e/ [https://perma.cc/S7CN-3BY7].

The White House has taken similar actions to encourage federal agencies to adopt AI. In February 2019, President Trump issued Executive Order 13859, Maintaining American Leadership in Artificial Intelligence, which requires agencies deploying AI to "pursue six strategic objectives in furtherance of both promoting and protecting American advancements in AI."[16] These objectives focus on research and development, interagency collaboration, education, and data and privacy protection.[17]

The following year, in December, President Trump issued Executive Order 13960, Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government.[18] This order requires federal agencies to adhere to nine principles "[w]hen designing, developing, acquiring, and using AI."[19] It also demands that agency AI be, among other things, accurate, lawful, secure from "adversarial manipulation," understandable, and "regularly monitored."[20] Under the order, humans are ultimately responsible for the technology and agencies must be transparent about AI usage.[21]

AI is not a partisan issue. President Biden has continued the work of his predecessor, launching the National Artificial Intelligence Research Task Force and AI.gov in accordance with Congress' National AI Initiative Act of 2020 to encourage AI use, not just in the federal government, but around the nation.[22]

### C. Implementation of AI by Agencies

The actions of elected leaders in Congress and the White House demonstrate the political willpower for the adoption and advancement of AI. This, however, has little practical impact unless administrative agencies are functionally applying the AI technology that the President and Congress seek to encourage. A recent report by members of ACUS indicates that agencies are doing just that.

---

[16] Exec. Order No. 13,859, 84 Fed. Reg. 3967, 3967 (Feb. 11, 2019).
[17] *See id.* at 3967–68.
[18] Exec. Order 13,960, 85 Fed. Reg. 78939, 78939 (Dec. 8, 2020).
[19] *Id.* at 78940.
[20] *See id.*
[21] *See id.*
[22] *The Biden Administration Launches the National Artificial Intelligence Research Task Force*, THE WHITE HOUSE (June 10, 2021), https://www.whitehouse.gov/ostp/news-updates/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/ [https://perma.cc/BXA5-Q3L6]; *The Biden Administration Launches AI.gov Aimed at Broadening Access to Federal Artificial Intelligence Innovation Efforts, Encouraging Innovators of Tomorrow*, THE WHITE HOUSE (May 5, 2021), https://www.whitehouse.gov/ostp/news-updates/2021/05/05/the-biden-administration-launches-ai-gov-aimed-at-broadening-access-to-federal-artificial-intelligence-innovation-efforts-encouraging-innovators-of-tomorrowz

After completing an assessment of 142 government agencies, ACUS members concluded that "[t]he use of AI-based tools to support government decision making, implementation, and interaction—what could be called 'algorithmic governance'—already spans the work of the modern administrative state."[23] Of those 142 agencies, sixty-four reported that they had at least experimented with AI.[24] Slightly more than half of the reported AI programs were coded in-house by the agency itself.[25] Agencies are using AI to perform every type of task that an agency may perform: enforcement, public engagement, regulatory research and analysis, and adjudication. [26] Algorithmic tools, ACUS reports, are "shaping, and in some cases displacing, the decisions of more senior agency decision-makers."[27]

## II. THE ROLE OF AI IN THE SOCIAL SECURITY DISABILITY CLAIM PROCESS

In 1972, Congress amended the Social Security Act to create the Supplemental Security Income Program under which those who are unable "to perform any substantial gainful activity by reason of a medically determinable impairment" can access disability benefits. [28] In 2020, nearly 1.6 million people applied for such benefits, down from 3.1 million in 2010. Roughly 600 thousand of the 1.6 million claimants in 2020 were awarded benefits.[29] Given the extraordinary number of applications, some claimants are required to wait years for a hearing. [30] Outcomes appear to vary dramatically from one decision-maker to the next—there are ALJs who award benefits over ninety percent of the time; others do so in less than ten percent of claims.[31]

---

[23] DAVID FREEMAN ENGSTROM ET AL., GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 7, 9 (2020). The authors of the report excluded military and intelligence agencies and agencies with fewer than four hundred employees.
[24] *See id.* at 7.
[25] *See id.* (reporting that 53% of the software programs were developed in-house).
[26] *See id.* at 17 ("agencies use AI to prioritize enforcement (e.g., prediction of potential violators of the federal securities laws at the SEC), engage with the public (e.g., a United States Citizenship and Immigration Services chatbot to provide assistance answering immigration questions), conduct regulatory research, analysis, and monitoring (e.g., Department of Health and Human Services tools to predict adverse drug events and unplanned hospital admissions), and adjudicate rights and benefits (e.g., United States Patent and Trademark Office tools to support patent and trademark determinations)").
[27] *Id.* at 11.
[28] ENGSTROM ET AL., *supra* note 23, at 38.
[29] *See id.* at 132.
[30] *See* ENGSTROM ET AL., *supra* note 23, at 38.
[31] *See id*.

To assist SSA decision-makers with this enormous case load and to somewhat standardize its decisions, the SSA has developed and deployed at least three different AI programs within the disability claim process: Quick Disability Determinations (QDD), the Insight Program, and the clustering of similar claims. Part II presents a brief overview of the disability claim process and explains how AI is used by the SSA to aid disability decision-makers.

### A. The Disability Claims Process

The disability claim process proceeds in five steps. In Step One, the agency determines whether the claimant is "doing substantial gainful activity."[32] In Step Two, the agency decides whether the claimant has "a severe medically determinable physical or mental impairment" or a severe combination of impairments.[33] The claim is denied if the claimant is either "doing substantial gainful activity" or is not severely impaired.[34] In Step Three, the agency evaluates whether the claimant's impairments "meet or equal one of" its listed impairments.[35] If they do, then the claimant is automatically found to be disabled.[36]

Before proceeding to Step Four, the agency assesses the claimant's residual functional capacity (RFC), which is "the most [a claimant] can still do despite [the person's] limitations."[37] It "is an assessment of an individual's ability to do sustained work-related physical and mental activities in a work setting on a regular and continuing basis."[38] This assessment is used in Step Four to determine whether the person is able to do his or her "past relevant work."[39] If not, the agency proceeds to the final step, Step Five, where it considers the claimant's RFC, "age, education, and work experience" to decide whether the individual is able to find "other work."[40] If, at the end of Steps Four and Five, the agency finds that the claimant is unable to either perform past work or transition to a new job, the person is awarded disability benefits.[41]

---

[32] Evaluation of Disability in General, 20 C.F.R. § 404.1520 (2022).
[33] *Id.*
[34] *See id.*
[35] *Id.*
[36] *See id.*
[37] Your Residual Functional Capacity, 20 C.F.R. § 404.1545 (2022).
[38] SOC. SEC. ADMIN., SOCIAL SECURITY RULING 96-8P, POLICY INTERPRETATION RULING: TITLES II AND XVI: ASSESSING RESIDUAL FUNCTIONAL CAPACITY IN INITIAL CLAIMS (1996).
[39] Evaluation of Disability, 20 C.F.R. § 404.1520 (2022).
[40] *Id.*
[41] *See id.*

There are four levels of review in a disability adjudication. The first two are at the state level when a state Disability Determination Service (DDS) reviews the individual's claim.[42] If the claimant disagrees with the first DDS determination, they can appeal for reconsideration by a separate examiner within the state's DDS.[43] If the DDS again denies the claim, the person can request a formal hearing by an ALJ, who will perform a de novo review.[44] If the ALJ reaches a finding of not disabled, the claimant can appeal for review by the SSA Appeals Council.[45] If, at the end of these four reviews, the SSA ultimately denies the claimant disability benefits, the individual can petition for judicial review in the federal court system.[46]

### B. The SSA's AI Tools

There are at least three ways that the SSA applies AI to assist in disability determinations: (1) predictive modeling software to identify those cases that should be easier to decide; (2) a natural language processing (NLP) program to evaluate ALJ opinions for inconsistencies; and (3) clustering algorithms to group similar claims together.

#### 1. Predictive Modeling

In 2008, the SSA began using predictive modeling software in the Quick Disability Determination (QDD) process to "identify cases where a favorable disability determination is highly likely and medical evidence is readily available."[47] The purpose of the program is to prioritize cases that are easier for decision-makers to resolve so as to "expedite case processing."[48] In 2010, the SSA issued a final rule formally adopting the QDD.[49] The SSA does not explain the operation of the AI software used in the QDD process in the

[42] *See* ENGSTROM ET AL., *supra* note 23, at 38.
[43] *See id.*
[44] *See id.*
[45] *See id.*
[46] *See id.*
[47] *Quick Disability Determinations (QDD)*, SOC. SEC. ADMIN., https://www.ssa.gov/disabilityresearch/qdd.htm#:~:text=The%20QDD%20process%20uses%20a,workload%20and%20expedite%20case%20processing (last visited Oct. 23, 2022) [https://perma.cc/TF5B-NX2K].
[48] *Id.*
[49] *See* Disability Determinations by State Agency Disability Examiners, 75 Fed. Reg. 62,676 (Oct. 13, 2010).

rule or elsewhere in the Federal Register, but it maintains on its website that engineers continue to work on improving the technology.[50]

In the QDD process, the SSA refers the 'easy' disability claims identified by the AI software to the state agency for review by "a designated disability determination examiner."[51] The disability determination examiner is given the discretion to consult state medical and psychological professionals before making a final decision on the claim.[52] If the state examiner denies the claim, or is unable to resolve a disagreement with the state medical or psychological consultants, then the claim must be adjudicated per the regular disability claim process.[53]

The SSA also uses predictive modeling at the hearing level to identify claims that "were denied reconsideration but have a high likelihood of receiving benefits."[54] These claims are then "moved ahead of cases in the queue with lower probabilities of allowance under the notion that disabled claimants should receive their decisions as soon as possible."[55]

### 2. Natural Language Processing

Insight is a software program that uses AI and natural language processing primarily "to flag potential policy compliance or internal consistency errors" in ALJ decisions.[56] It is used both by ALJs at the hearing level—to review the drafts of their opinions for inconsistencies—and by the Appeals Council—to identify potential errors in the underlying ALJ's decision.[57] Insight is mandatory for fully favorable decisions by ALJs, but it is voluntary for both unfavorable decisions by ALJs and the Appeals Council.[58] In 2019, however, the Inspector General recommended, and the SSA agreed, that the Appeals Council should be required to use Insight.[59]

---

[50] *See Quick Disability Determinations (QDD)*, *supra* note 47 ("We continue to refine the QDD predictive model to reflect the characteristics of the recent applicant population and optimize its ability to identify strong candidates for expedited processing.").
[51] Quick Disability Determination Process, 20 CFR § 404.1619 (2022).
[52] *See id.*
[53] *See id.* By requiring that unfavorable claims be adjudicated through the regular disability process, the SSA all but ensures that its QDD software will never be the subject of judicial review. A triumphant claimant would have no standing to petition for review of a favorable decision by the QDD AI.
[54] ENGSTROM ET AL., *supra* note 23, at 40.
[55] *Id.*
[56] OFFICE OF THE INSPECTOR GEN., SSA'S USE OF INSIGHT SOFTWARE TO IDENTIFY POTENTIAL ANOMALIES IN HEARING DECISIONS 1 (2019), https://oig-files.ssa.gov/audits/full/A-12-18-50353.pdf [https://perma.cc/ZZT5-SYVD].
[57] *See* ENGSTROM ET AL., *supra* note 23, at 40.
[58] *See id.*; OFFICE OF THE INSPECTOR GEN., *supra* note 56, at 11.
[59] *Id.* at E-2.

Insight functions by reviewing an ALJ's draft or final decision and flagging issues that "are suggestive of policy noncompliance or internal inconsistencies."[60] This review is no cursory spellcheck. There are over 30 "quality flags," as they are called, for which Insight is searching. Some of these flags are highly sophisticated. For example, the most frequently triggered "quality flag" is: "The decision contains language describing what the claimant cannot do rather than describing the claimant's maximum capacity to perform work-related physical or mental functions."[61] Another flag, the fourth most activated, warns: "The decision is vague about whether an impairment is relevant to the period at issue or whether it is an impairment at all."[62]

Insight is not only capable of reviewing decisions, but it also assists in writing them, functioning as an "analysis template generator." In performing this task, it scans a case file; extracts key information such as disposition, disposition date, and claimant's height and weight; and uses that information to calculate values such as the claimant's body mass index and filing deadline.[63] The program then populates a template with information about the case to save decision-makers from having to manually enter the data.[64]

### 3. Clustering

The third type of AI used by the SSA is clustering algorithms used to group similar claims together.[65] These coalesced claims are then assigned to the same adjudicator who, through repetition and familiarity, becomes better at deciding that specific type of claim.[66] The purpose of clustering algorithms is to help an adjudicator develop a "micro-specialization" for a particular type of disability claim.[67]

## III. JUDICIAL REVIEW OF SOCIAL SECURITY DETERMINATIONS AND THE "SUBSTANTIAL EVIDENCE" STANDARD

---

[60] ENGSTROM ET AL., *supra* note 23, at 40.
[61] OFFICE OF THE INSPECTOR GEN., *supra* note 56, at 7.
[62] *Id.*
[63] *See* ENGSTROM ET AL., *supra* note 23, at 40 ("using information extraction from the case management system, Insight provides a case summary, including claim type, disposition, claimant information (date of birth, claimed onset dates, body mass index), and acquiescence rulings for the region"); OFFICE OF THE INSPECTOR GEN., *supra* note 57, at B-1 to B-2.
[64] OFFICE OF THE INSPECTOR GEN., *supra* note 56, at B-2.
[65] *See* ENGSTROM ET AL., *supra* note 23, at 39.
[66] *See* Engstrom & Ho, *supra* note 2, at 811.
[67] *See* ENGSTROM ET AL., *supra* note 23, at 39.

If the Appeals Council upholds an ALJ's unfavorable disability determination, the claimant has the opportunity to petition for judicial review of the SSA's decision in federal court. The court will affirm the ALJ's decision as long as it is supported by "substantial evidence."[68] In general, courts have interpreted the "substantial evidence" standard to require them to evaluate the merits of the ALJ's decision-making *process*, rather than the final decision.[69] This is problematic for factual determinations made by AI whose decision-making process is, for many reasons, inscrutable.[70] Part III lays out the substantial evidence standard as it is applied today and then explains why it is unsuitable for AI decisions.

### A. The Current "Substantial Evidence" Standard

To develop an understanding of the "substantial evidence" standard, one must begin, as always, with the statutory language. The "substantial evidence" standard is prescribed in both the APA and the Social Security Act. For agencies in general, the APA requires that decisions derived through formal hearings conducted "on the record" be "set aside" when they are "unsupported by substantial evidence."[71] For the SSA in particular, the Social Security Act states that the agency's factual findings are conclusive "if supported by substantial evidence."[72] The "substantial evidence" standard in the APA and the Social Security Act require the same "quantum of factual support" and, for all intents and purposes, are the same.[73]

These Congressional standards are somewhat vague and ambiguous. Congress does not specify what counts as "evidence," or how much is needed to surpass the "substantial" benchmark. Congress does little more than provide the words "substantial evidence." Therefore, to cultivate an understanding of how the "substantial evidence" standard might be applied to a specific scenario, one must turn to the federal judiciary and consider the development

---

[68] *See* 42 U.S.C. § 405(g); 5 U.S.C. § 706(2)(e).

[69] *See Allentown Mack Sales & Serv*., 522 U.S. at 374 ("the process by which [the agency] reaches the result must be logical and rational").

[70] *See* Nicholas Berente et al., *Managing Artificial Intelligence*, 45 MIS QUARTERLY 1433, 1443–44 (2021) ("whether one can explain, interpret, or understand AI decisions depends on the algorithm and its opacity and explainability, as well as the transparency decisions, and interpretation of humans").

[71] 5 U.S.C. § 706(2)(e) ("The reviewing court shall hold unlawful and set aside agency action, findings, and conclusions found to be unsupported by substantial evidence . . .").

[72] 42 U.S.C. §405(g) ("The findings of the Commissioner of Social Security as to any fact, if supported by substantial evidence, shall be conclusive.").

[73] Ass'n of Data Processing Serv. Orgs., Inc. v. Bd. of Governors of Fed. Res. Sys., 745 F.2d 677, 686 (D.C. Cir. 1984) ("the § 1848 'substantial evidence' requirement applicable to our review here demands a quantum of factual support no different from that demanded by the substantial evidence provision of the APA").

of the standard by the Supreme Court and its application by the federal courts.[74]

In applying a statutory standard of review, courts attempt to identify and apply the "mood" expressed by Congress in the legislation.[75] This mood is intended to "serve as a standard for judgment and not as a body of rigid rules assuring sameness of application." [76] The "mood" expressed by the "substantial evidence" standard is one of respect towards the agency fact-finder.[77] In applying the standard, a court "gives the agency the benefit of the doubt" by asking not whether the agency's conclusion is required by the evidence, but instead whether the agency's conclusion is *reasonable* based on the evidence on the record.[78]

The Supreme Court conveys this deferential mood by specifying that the amount of evidence required "is not high" and by explaining that "substantial" actually means "sufficient" or "more than a mere scintilla."[79] In applying the standard, the Supreme Court looks for "such relevant evidence as a reasonable mind might accept as adequate to support a conclusion."[80] Agencies also cannot cherry-pick evidence, "but must draw all those inferences that the evidence fairly demands."[81]

Decisions from the district courts and courts of appeals, which enforce the standard that the Supreme Court reads into the APA and the Social Security Act, have common themes. In general, courts find that a disability decision is unsupported by substantial evidence when the ALJ ignores self-imposed SSA regulations, fails to address why he or she rejected contrary evidence, or does not explain how the conclusion was derived from the evidence on the record, sometimes called building "an accurate and logical bridge."[82] In coming to a

---

[74] *See* Kristin Hickman & R. David Hahn, *Categorizing Chevron*, 81 OHIO ST. L. J. 612, 655 (2020) ("[T]he statutory requirement is naught but the label . . . The real meat of any statutory standard of review lies in the judicially-developed boilerplate and operational details. Thus, even with a statutory standard of review, courts have a fair degree of latitude in adjusting over time the details that operationalize the standard.").

[75] *See* Universal Camera Corp. v. NLRB, 340 U.S. 474, 487 (1951).

[76] *Id*.

[77] *See id*. at 490 ("The Board's findings are entitled to respect . . .").

[78] Allentown Mack Sales & Serv. v. NLRB, 522 U.S. 359, 377 (1998).

[79] Biestek v. Berryhill, 139 S. Ct. 1148, 1154 (2019).

[80] *Id.* (quoting Consolidated Edison Co. v. NLRB, 305 U.S. 197, 229 (1938)).

[81] *Allentown Mack Sales & Serv.*, 522 U.S. at 378.

[82] *See, e.g.* Shinaberry v. Saul, 952 F.3d 113, 123 (4th Cir. 2020) (explaining that the ALJ was required to give the claimant's treating physician controlling weight because it was required by 404.1527(c)(2)); Hargett v. Comm'r of Soc. Sec., 964 F.3d 546, 551 (6th Cir. 2020) ("even where the ALJ's findings are otherwise supported by substantial evidence, the ALJ's failure to follow agency rules or regulations is a ground for reversal."); Deborah M. v. Saul, 994 F.3d 785, 788 (7th Cir. 2021) ("[a]lthough the ALJ need not discuss every piece of evidence in the record, he must confront the evidence that does not support his conclusion

decision, courts review "the entire record."[83] ALJs may be permitted to discredit a claimant's testimony for lack of credibility.[84]

What this means for ALJs is that they must show their work. They must provide an in-depth detailed discussion of the facts, explaining how they reached their decision and why they discounted evidence that goes against their conclusion. If they discount testimony of subjective symptoms by the claimant, then they must conduct additional analysis explaining the basis for their credibility determination. Overall, the ALJ must provide evidence of "reasoned decisionmaking" and demonstrate that the process by which they reached their decision was "logical and rational."[85]

### B. Inadequacy of the "Substantial Evidence" Standard for Reviewing Agency AI-Driven Decisions

AI raises myriad problems. Some highlighted by the ACUS in its statement on agency use of AI include: AI can reflect or intensify the biases of its programmers or the historical data used to "teach" the software; AI software requires a large amount of data, the mining of which can create privacy issues; AI programs may be susceptible to hacking or manipulation by bad actors; and AI can distort an agency's approved discretional hierarchy when appointed agency decision-makers come to reflexively rely on AI output, thus transferring practical authority to software engineers.[86] All of these challenges merit significant discussion; however, the challenge that bears down specifically upon the "substantial evidence" judicial standard is the inscrutability of AI-derived decisions.

AI is inscrutable because its reasoning is opaque, nonintuitive, complex, and not readily transparent.[87] Simple AI technology, which is

---

and explain why it was rejected.") (internal quotations omitted); Craft v. Astrue, 539 F.3d 668, 673 (7th Cir. 2008) ("[An ALJ] must provide an 'accurate and logical bridge' between the evidence and the conclusion that the claimant is not disabled, so that 'as a reviewing court, we may assess the validity of the agency's ultimate findings and afford [the] claimant meaningful judicial review.'") (quoting Young v. Barnhart, 362 F.3d 995, 1002 (7th Cir. 2004).

[83] *See* Ahearn v. Saul, 988 F.3d 1111, 1115 (9th Cir. 2021).

[84] *See* Thomas v. Barnhart, 278 F.3d 947, 958–59 (9th Cir. 2002) (explaining that courts may consider "[claimant's] reputation for truthfulness, inconsistencies either in [claimant's] testimony or between [her] testimony and [her] conduct, [claimant's] daily activities, [her] work record, and testimony from physicians and third parties concerning the nature, severity, and effects of the symptoms of which [claimant] complains.").

[85] *See Allentown Mack Sales & Serv.*, 522 U.S. at 374.

[86] *See* Statement #20, *supra* note 2, at 6617–18.

[87] *See* Berente et al., *supra* note 70, at 1441 (describing the four causes of AI inscrutability as opacity, transparency, explainability, and interpretability); Danaher et al., *Algorithmic Governance: Developing a Research Agenda Through the Power of Collective Intelligence*,

relatively easy to understand and predict, is primarily rule-based, "explicit logic coded into technology."[88] It consists mostly of "if . . . then" statements. Such AI software would ace the logic games portion of the LSAT. More advanced AI software, on the other hand, has built-in learning capabilities. It is designed to replicate the neural reasoning of the human brain by using multiple layers of nodes, "which transfer information to each other and learn on their own how to weigh connections between nodes."[89] Rather than being deterministic, like a simple rule-based algorithm, it is probabilistic.[90]

Because AI has the ability to learn, the algorithm can become opaque, even to the software engineers who designed the program.[91] "[B]ecause a machine learning system learns on its own and adjusts its parameters in ways its programmers do not specifically dictate, it often remains unclear precisely how the system reaches its predictions or recommendations."[92] Even when it is clear that an algorithm considered, for example, a claimant's subjective symptoms or a treating physician's opinion, it may be difficult to identify how the AI program weighed such evidence or whether those factors influenced its decision.[93] This is a particular challenge under the "substantial evidence" standard when judges expect ALJs to assign more weight to certain evidence and testimony.[94]

Even when software engineers are able to understand and describe the AI program's reasoning process, it may be of little help to judges reviewing the decision if the reasoning is nonintuitive.[95] The AI program's ultimate determination may be reasonable, but the reasoning applied by the AI might

---

BIG DATA & SOCIETY, July–Dec. 2017, at 3 (explaining that opacity and transparency create problems for algorithmic governance); Engstrom & Ho, *supra* note 2, at 824–25 (calling AI inscrutable and nonintuitive).

[88] *See* Berente et al., *supra* note 70, at 1441.

[89] Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 COLUM. L. REV. 1829, 1832–33 (2019).

[90] *See* Berente et al., *supra* note 70, at 1441.

[91] *See* ENGSTROM ET AL., *supra* note 23, at 11 ("[E]ven a system's engineers may not fully understand how it arrived at a result."); Berente et al., *supra* note 67, at 1441 ("The logic of some advanced algorithms is simply not accessible, even to the developers of those algorithms.").

[92] *See* Deeks, *supra* note 89, at 1832.

[93] ENGSTROM ET AL., *supra* note 23, at 11 ("[S]tate of the art machine learning deploys far more complex models to learn about the relationship across hundreds or even thousands of variables. Model complexity can make it difficult to isolate the contribution of any particular variable to the result.").

[94] *See, e.g.*, Ford v. Saul, 950 F.3d 1141, 1154 (2020) ("As a general rule, a treating physician's opinion is entitled to 'substantial weight.'").

[95] ENGSTROM ET AL., *supra* note 23, at 11 ("[AI programs] operate according to rules that are so complex, multi-faceted, and interrelated that they defy practical inspection, do not comport with any practical human belief about how the world works, or simply lie beyond human-scale reasoning.").

be illogical or outside the confines of what human decision-makers are supposed to consider in that situation.[96] This type of scenario could arise if the AI software detects a pattern in human decision-making that we ourselves do not recognize. The AI decision-maker reached the right decision but for the "wrong" reasons. This nonintuitive nature of AI is problematic for judges because the reasoning and decision-making process is the focus of the "substantial evidence" standard, not the adequacy of the final decision.

Even when software engineers are able to understand and describe the AI program's reasoning process *and* the AI's reasoning is intuitive; lawyers, judges, and claimants lacking experience in the subject area may not be able to comprehend the software's operation. Lawyers, as officers of the court, are charged with aiding the judge in arriving at the proper decision. If a person requires a doctorate in software engineering to understand the algorithm, then most lawyers will be poorly suited for arguing their positions to the judge and helping him or her to arrive at the best decision. Similarly, it may be difficult for a judge to properly assess the AI decision without relying on help from an outside technical expert. Finally, when the judge writes the opinion, it must be done in a manner that the claimant, the primary target for the opinion, can understand. An opinion filled with confusing technical language that the judge only understood after explanation from a software engineer would be unlikely to assure a claimant that the case was resolved properly.

Finally, there is the barrier of transparency, that is "an openness or willingness" on the part of the software's owner to disclose the technology.[97] Even if all of the above could be satisfied, if the AI decision-making process was clear, reasonable, and understandable, the agency or the software developer may not wish to disclose the inner workings of the AI program.

There are two main reasons that an agency or software developer may offer for maintaining a lack of transparency. First, revealing information about the algorithm could enable "gaming or adversarial learning" by future claimants.[98] For example, if a doctor or claimant learns what boilerplate language influences the software, they may manipulate the medical records to include such language. Second, if the software is owned by a third party, it "may be protected by the same patent, copyright, or trade secrecy rules that apply in the private sector context."[99] To avoid the latter issue, the SSA should ensure that its software contractors have no such secrecy requirements.

---

[96] *See id*. ("Even if data scientists can spell out the embedded rule, such rules may not tell a coherent story about the world as humans understand it, defeating conventional modes of explanation.").

[97] Berente et al., *supra* note 70, at 1441.

[98] Statement #20, *supra* note 2, at 6617.

[99] Engstrom & Ho, *supra* note 2, at 841. *See also* Statement #20, *supra* note 2, at 6617 ("When agencies' AI systems rely on proprietary technologies or algorithms the agencies do

The "substantial evidence" standard, as it is currently applied, is poorly suited for evaluating inscrutable AI decisions. Unless Congress acts to upgrade the APA and make it more amenable to agency AI, the federal judiciary will be required to revise the standard on its own. Part IV suggests several factors that the courts may consider in making such a revision.

## THE "SUBSTANTIAL EVIDENCE FOR AI" STANDARD

The "substantial evidence for AI" standard should require courts to ask two key questions. First, is the AI reasonable? Second, is the decision based on "such relevant evidence as" a reasonable AI "might accept as adequate to support a conclusion"?[100] The inscrutability of AI precludes courts from closely examining the decision-making process. Therefore, the court must ensure that the AI decision-maker was functioning reasonably, and that the data was sufficient to enable that reasonable AI to come to a reasonable conclusion.

### A. Is the AI Software Reasonable?

When performing judicial review of disability determinations, courts do not ask, "is the ALJ reasonable?"[101] Rather, courts question whether the decision-making process was reasonable. Because AI is nothing more than a computerized decision-making process, evaluating the reasonableness of the AI itself is functionally equivalent to asking whether an ALJ's decision-making process is reasonable.

There are a number of factors for courts to consider when questioning the reasonableness of an AI program: the expertise of the AI's human developers, the adequacy of the training dataset, and the oversight protections in place to prevent the development of harmful biases.

First, to ensure the sufficiency of the baseline algorithm, reasonable AI should be created by people with the appropriate knowledge. The computer engineers, or development team, should have both a technical understanding

not own, the agencies and the public may have limited access to the information about the AI techniques.").

[100] Recall the Supreme Court's definition of "substantial evidence" as "such relevant evidence as a reasonable mind might accept as adequate to support a conclusion." Biestek v. Berryhill, 139 S. Ct. 1148, 1154 (2019) (quoting Consolidated Edison Co. v. NLRB, 305 U.S. 197, 229 (1938)).

[101] The propriety of ALJs is the subject of a different line of case law. *See generally* Lucia v. SEC, 138 S. Ct. 2044 (2018) (considering whether SEC ALJs were officers for purposes of the Appointments Clause); Freytag v. Comm'r, 501 U.S. 868 (1996) (considering whether Special Tax Judges were constitutionally appointed).

of AI software and an in-depth understanding of the disability process.[102] The Insight Program, for example, was created by Kurt Glaze, a software engineer who spent years as an adjudicator for the SSA.[103] This requirement need not be satisfied by a single individual, like Mr. Glaze, but could be met when the team that created the AI included representatives from both the SSA and the computer science field.

Second, to ensure the sufficiency of the algorithm's growth, reasonable AI must have been exposed to an adequate training dataset.[104] Most contemporary AI software programs use learning algorithms. They grow and develop as they are exposed to more and more data. Their performance, therefore, depends significantly on the scope and quality of the training data. Agency AI must be exposed to a sufficient amount of the appropriate data before it can be used to adjudicate factual determinations.

In assessing a dataset, courts should rely upon the framework used by computer scientists and engineers, called "The Five V's of Big Data." These are: volume, variety, velocity, value, and variability.[105] Volume refers to the amount of data.[106] Variety defines the diversity of the dataset.[107] Velocity signifies the rate at which data can be received and processed.[108] Value suggests that the data must be applicable to the specific objective of the algorithm.[109] Variability refers to the extent to which the same variable in a dataset has multiple meanings.[110]

Applying "The Five V's of Big Data," a court reviewing an algorithm's training data should require, to satisfy volume and variety, that it include a large and diverse amount of healthcare information from a similarly large and diverse population of human subjects. In the spirit of velocity, the training

---

[102] *See* ENGSTROM ET AL., *supra* note 23, at 44 ("[T]he SSA experience illustrates the importance of blending subject-matter and technical expertise.").
[103] *See id.*
[104] Training data refers to the "initial set of data used to help a program understand how to apply technologies like neural networks to learn and produce sophisticated results." *Training Data*, TECHOPEDIA (Feb. 17, 2022), https://www.techopedia.com/definition/33181/training-data [https://perma.cc/H5Y9-WBUC].
[105] *See* Judith J. Warren, *A Big Data Primer*, in BIG DATA-ENABLED NURSING 33, 37 (Connie W. Delaney et al. eds., 2017).
[106] *See* Harry Hemingway et al., *Big Data From Electronic Health Records for Early and Late Translational Cardiovascular Research: Challenges and Potential*, 39 EUR. HEART J. 1481, 1482 (2018).
[107] *See* Anil Jain, *The 5 V's of Big Data*, IBM: WATSON HEALTH PERSPECTIVES (Sept. 17, 2016), https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/ [https://perma.cc/M35N-E73Y].
[108] *See id.*
[109] *See id.*
[110] *See id.* For example, a cardiologist might use "CP" to mean "chest pain" while a neurologist would use "CP" to denote "cerebral palsy."

dataset should be continuously updated to account for changing conditions that might affect human health, like a global pandemic. To have value, the data should be applicable to disability determinations. It should consider medical information and other factors such as "physical environment, consumer information, socioeconomic and behavioral [sic] factors and user-generated data from mobile health apps, wearables, sensors and social media."[111] Finally, the court should verify that there is little variability across the dataset, or, at least, that the algorithm accounts for possible variability that may result from a lack of standardization across the medical field. As AI begins to play a larger role in healthcare, it is conceivable that doctors will adjust to standardize their medical records and nomenclature to make it easier for AI to learn and adapt.

The third and final requirement for reasonable AI should be that the human managers be able to demonstrate that they have accounted for and taken action to prevent the existence and exacerbation of harmful bias. By removing a subjective human mind from the decision-making process, AI is able to avoid implicit bias. However, if the software developers are biased or the training dataset contains biases, then the AI could "increase [that bias] over time through feedback."[112] This can occur when the AI makes biased determinations "which are then reflected in the data set or data environment the system uses to make future predictions."[113] To prevent biases, the AI management team must periodically monitor the algorithm for such bias and frequently upgrade the AI to ensure it stays up to date on "[d]ata science techniques for identifying and mitigating harmful biases."[114] Courts should require agencies to demonstrate this internal surveillance in order to prove that the AI is reasonable.

### B. Is the Decisions Based on Such Relevant Evidence as a Reasonable AI Might Accept as Adequate to Support a Conclusion?

Having evaluated whether the AI itself was reasonable, courts should next turn to the evidence used to make the particular decision. Courts should evaluate whether the AI's decision was based on such relevant evidence as a reasonable AI might accept as adequate to support a conclusion. Once again, courts should rely upon the format provided by "The Five V's of Big Data." Four of the five principles—volume, variety, value, and variability—are applicable to individual determinations.

---

[111] Hemingway et al., *supra* note 106, at 1482.
[112] Statement #20, *supra* note 2, at 6616.
[113] *Id.*
[114] *Id.*

To ensure an appropriate volume, the court should ensure that there is a sufficient amount of information about the claimant and that there are no significant gaps or missing data. To confirm adequate variety, courts should require that the dataset consist of both medical and nonmedical evidence. The medical evidence should be diverse on its own. This can be determined by asking the question: is the medical evidence derived only from observations at a physical exam, or does it also contain information from x-rays, blood tests, and body imaging scans? In considering value, courts should ensure that the input data is applicable to the alleged disability. Dental exams, for example, would be inadequate to support a conclusion regarding an alleged foot injury. Finally, the court should review the variability of the data to assess whether a lack of standardized medical reports might have led the AI to make an inappropriate decision.

## CONCLUSION

Administrative agencies' factual determinations derived during formal adjudications are upheld by the federal judiciary as long as administrators demonstrate "reasoned decisionmaking" that is supported by "substantial evidence."[115] This deference to agency decisions has led some to argue that administrative law provides the perfect forum for applying and evaluating AI decisions since courts do not "demand anything close to transparency."[116] While there is a certain level of deference programmed into administrative law, the APA still requires "meaningful review."[117] Courts are "not simply rubber-stamping agency fact-find[ers]."[118] The federal judiciary's current standard for "meaningful review," however, is ill-suited for resolving administrative agency AI decisions because the AI decision-making process is largely inscrutable.

Ideally, Congress would amend the APA to provide specific stipulations to agencies using AI and to courts reviewing AI decisions. In the absence of congressional action, however, the federal judiciary must carry on as it has for the last seventy-six years by updating and revising the APA's judicial standard to meet the exigencies of the current time.

An effective "substantial evidence for AI" standard would ask two questions. First, is the AI reasonable? Second, is the AI's decision based on

---

[115] *See, e.g.*, Biestek v. Berryhill, 139 S. Ct. 1148, 1154 (2019) ("The phrase 'substantial evidence' is a 'term of art' used through administrative law to describe how courts are to review agency factfinding."); *Allentown Mack Sales & Serv.*, 522 U.S. at 374 ("adjudication is subject to the requirement of reasoned decisionmaking as well").

[116] Deeks, *supra* note 89, at 1840.

[117] Dickinson v. Zurko, 527 U.S. 150, 162 (1999).

[118] *Id.*

such relevant evidence as a reasonable AI might accept as adequate to support a conclusion? If a court affirmatively answers both questions, then the agency's decision should be affirmed. These two questions working together can guarantee that the court has met the APA's requirement for "meaningful review" without adopting a more scrutinizing "mood" than the deferential "substantial evidence" standard calls for.

The day is coming when the courts will be called upon to address an agency AI decision. Given the vast amount of disability determinations and the SSA's use of AI, it is likely that the collision between the "substantial evidence" standard and AI will occur on a petition for judicial review of an unfavorable disability determination. The courts must keep a weathered eye on the horizon and be prepared for the moment when it arrives.