VAVILOV AND GENERATIVE AI

ROBERT DIAB*

This article considers whether a decision made by generative artificial intelligence can satisfy the standard of reasonableness set out in Canada (Minister of Citizenship and Immigration) v. Vavilov. Vavilov *requires that administrative decisions be justified through reasons that are transparent and intelligible to the affected party. Earlier scholarship, law, and policy have assumed that AI cannot do this because it cannot provide reasons and its inner workings are opaque or uninterpretable. However, new capabilities of large language models challenge this view. Recent experiments show that when prompted with party submissions and relevant legal materials, generative AI can produce persuasive, legally grounded reasons for decisions. The article evaluates two responses: one argues that AI decisions remain unreasonable under* Vavilov *since their true basis lies in opaque technical processes; the other contends that* Vavilov *focuses on the cogency of stated reasons, not how they were generated. The article supports the latter position, suggesting that* Vavilov *leaves open the possibility that AI-generated decisions can be reasonable, provided their reasons meet the decision-making standard applied to human actors.*

INTRODUCTION

As debate over the impact of the Supreme Court of Canada's decision in *Canada (Minister of Citizenship and Immigration) v. Vavilov* continues to unfold, developments in generative artificial intelligence (AI) have given rise to an important question on which *Vavilov* is central but silent.[1] *Vavilov* clarifies how the standard of reasonableness should apply to administrative decisions on judicial review. In broad terms, reasons given in support of a decision must "justify to the affected party, in a manner that is transparent and intelligible, the basis on which [a decision-maker has] arrived at a particular conclusion."[2] Scholars and policy-makers have assumed that AI could not be used to make or assist in making high-

[*] Professor, Faculty of Law, Thompson Rivers University.

[1] 2019 SCC 65 [*Vavilov*].

[2] *Ibid* at para 96; see also *Dunsmuir v New Brunswick*, 2008 SCC 9 at para 47 [*Dunsmuir*].

impact administrative decisions because — among other potential issues — it could not justify the basis of them in a "transparent and intelligible" manner.

Most of this commentary, however, predates the advent of generative AI. It is focused on forms of AI that could calculate scores or percentages (for example, risks of recidivism or sentencing ranges), but could not provide reasons.[3] Many of these earlier forms of AI raised concerns about output resting on machine learning processes that are mostly opaque, concealing bias and error.[4] AI should not be used to decide high-impact cases, many argued, because its use would not be procedurally fair or reasonable — due to bias, lack of notice, or the failure to give adequate reasons, among other grounds.[5]

But large language models have brought about a new possibility only now being explored — one that complicates assumptions about AI in relation to fairness and reasonableness. Experiments by scholars and jurists are beginning to show how language models can be used to provide a cogent and persuasive justification for a decision, one that is responsive to party submissions and applies relevant law to the facts in a case correctly.[6] The key to doing so is to upload to the model materials submitted by the parties and to prompt the model to apply a discrete body of law to those materials. The reader is invited to consult sources canvassing these experiments in detail,[7] but to assume for the purpose of this article that language models are capable of producing reasons to justify an administrative decision comparable in quality to those of a human decision-maker. A host of questions arise from this possibility, including whether full reliance on AI in this manner would be perceived to be legitimate or whether significant reliance on AI to assist in making a decision (for example, having AI draft a decision and merely affirming it) would involve a form of subdelegation.[8] The purpose of this article is to ask, more specifically, whether a decision by AI alone could withstand review on the reasonableness standard as contemplated in *Vavilov*.

As noted above, the key requirement in *Vavilov* is that reasons provided to support a decision "justify to the affected party, in a manner that is transparent and intelligible, the basis on which [the decision-maker] arrived at a particular conclusion."[9] The Supreme Court in *Vavilov* assumes that where written reasons purport to explain how a decision was reached,

[3] See e.g. Amy Goudge, "Administrative Law, Artificial Intelligence, and Procedural Rights" (2021) 42 Windsor Rev Legal Soc Issues 17 at 24–25; Jennifer Raso, "AI and Administrative Law" in Florian Martin-Bariteau & Teresa Scassa, eds, *Artificial Intelligence and the Law in Canada* (Toronto: LexisNexis Canada, 2021), ch 7; Jesse Beatson, "AI-Supported Adjudicators: Should Artificial Intelligence Have a Role in Tribunal Adjudication?" (2018) 31:3 Can J Admin L & Prac 307; Richard M Re & Alicia Solow-Niederman, "Developing Artificially Intelligent Justice" (2019) 22:2 Stan Tech L Rev 242.

[4] Sources for this are noted in Part I below.

[5] Sources canvassed in Part I below include Goudge, Raso, Re & Solow-Niederman, *supra* note 3.

[6] Adam Unikowsky, "In AI We Trust", *Adam's Legal Newsletter* (8 June 2024), online: [perma.cc/U648-PDSE] [Unikowsky, "Trust"]; Adam Unikowsky, "In AI We Trust, Part II", *Adam's Legal Newsletter* (June 16, 2024), online: [perma.cc/D3GN-4XUL] [Unikowsky, "Trust Part II"]; Robert Diab, "The Evolving Role of AI in Legal Judgment" *Law, Innovation, and Technology* [forthcoming in 2026]; Paul Daly, "Artificial Intelligence and Administrative Tribunals" in Yee-Fee Ng & Matthew Groves, eds, *Automation in Public Governance: Theory, Practice and Problems* (Hart, Oxford, 2025) at 16–19 [Daly, "Artificial Intelligence"].

[7] *Ibid*.

[8] For an overview of these issues, see Daly, "Artificial Intelligence," *supra* note 6; Paul Daly, "Artificial Administration: Administrative Law, Administrative Justice and Accountability in the Age of Machines" (2023) 30:2 Austl J Admin L 95 [Daly, "Artificial Administration"].

[9] *Vavilov*, *supra* note 1 at para 96.

those reasons provide the basis for the decision. The Supreme Court is silent on the possibility that generative AI brings about: providing cogent reasons that offer an intelligible, *rational* basis for arriving at a certain outcome, while the actual, *technical* basis for reaching it involves something different (such as correlations and algorithms) and is mostly opaque. Should a court on judicial review, assessing reasonableness, accept a language model's textual reasons as the *basis* for that decision? If not, can it be reasonable?

This article canvasses two opposing positions on these questions, and it supports one over the other as more consistent with *Vavilov*. Skeptics contend that even if AI can produce a decision with reasons that make its rationale transparent and intelligible, it cannot be reasonable under *Vavilov*.[10] A model's true basis for decision can only be statistical rather than normative, and it cannot be rendered free of bias or error to a reasonable degree.[11] The opposing view contends that *Vavilov*'s emphasis on the rationale set out in written reasons makes it possible for language models to meet the standard of reasonableness if certain conditions are met.[12] Concerns about bias or opacity in a model's inner workings can be addressed by fine-tuning, auditing, and testing, which would form the basis of a presumption of impartiality, as is the case with human decisions.[13] The test should be whether a model's reasons are cogent and responsive, not whether the model's underlying process is statistical or fully understood.

Part I of this article sets out a brief overview of earlier assumptions about AI's functional limitations and new capabilities language models bring about. Part II canvasses the Supreme Court's clarification of the reasonableness standard in *Vavilov* in light of the use of AI as contemplated in recent experiments. It then considers arguments for and against an AI decision withstanding a challenge for reasonableness.

I. EARLIER ASSUMPTIONS AND NEW CAPABILITIES OF AI

To lend context to the discussion of AI and *Vavilov*, it would be helpful to first address how new capacities in AI challenge earlier assumptions about AI's role in administrative judgment — in part to explain why it was not on the horizon of the Supreme Court's otherwise comprehensive discussion of reasonableness in *Vavilov*.

A. EARLIER ASSUMPTIONS IN SCHOLARSHIP AND POLICY

[10] I assemble this argument drawing on a host of sources including: Amin Ebrahimi Afrouzi, "John Robots, Thurgood Martian, and the Syntax Monster: A New Argument Against AI Judges" (2024) 37:2 Can JL & Jur 369; John Tasioulas, "The Rule of Algorithm and the Rule of Law" (2023) in Christoph Bezemek, Michael Potacs & Aleander Somek, eds, *Vienna Lectures on Legal Philosophy, Volume 3: Legal Reasoning* (Oxford: Hart Publishing, 2023) 17; Goudge, *supra* note 3; Re & Solow-Neiderman, *supra* note 3.

[11] Afrouzi, *supra* note 10; Tasioulas, *supra* note 10. Both sources are discussed in more detail in Part II below.

[12] For this argument, I draw on Eugene Volokh, "Chief Justice Robots" (2019) 68:6 Duke LJ 1135; Richard Susskind, *Online Courts and the Future of Justice* (Oxford: Oxford University Press, 2019), ch 27.

[13] Sources on these technical possibilities are canvassed in Part II below; on the presumption of impartiality, see especially *Keita v Canada (Citizenship and Immigration)*, 2015 FC 1115 at para 1 [*Keita*]; see also *Panov v Canada (Minister of Citizenship and Immigration)*, 2015 FC 716 at para 20 [*Panov*]; *Arthur v Canada (Attorney General)*, 2001 FCA 223 at para 8 [*Arthur*].

Much of the commentary on AI in administrative law predates the release of ChatGPT in the fall of 2022.[14] A common element in the literature is the focus on forms of AI that output scores or probabilities based on similarities or pattern-matching. Common examples include software for predicting recidivism (used by parole boards), deciding immigration eligibility, detecting tax evasion or welfare fraud, and allocating social benefits.[15] In some cases, the code or algorithms underlying these forms of AI are static and follow a clear set of rules that can be readily understood and examined for bias.[16] But much of AI involves machine learning processes that are so complex as to render the bases on which a program generates a specific output "inherently opaque."[17] Developers and auditors of the software alike may come to know broadly "how data moves through each layer of the network," but not "the specifics."[18] Forms of AI involving several layers of machine learning — including language models — are assumed to be explainable in some measure but not interpretable.[19] This makes it difficult if not impossible to discern precisely how a model has made a decision in a given case, that is, "the factors it weighs and the correlations it draws."[20]

Scholars have highlighted three problems with using forms of AI involving machine learning to make administrative decisions. One is that AI could not itself provide reasons to justify or explain a given outcome, since the basis of the decision remains hidden behind opaque algorithmic processes.[21] A second, related problem is that because AI tools conceal the technical basis or process by which they arrive at a decision, it is impossible to ascertain whether the process entailed bias or was otherwise improper. As one scholar writes: "Without being able to identify the factors that the AI considered through a process of autonomous self-learning, it would be impossible to communicate the reasons for the AI's outcome or to ensure the AI's reasoning process accounted for all relevant factors."[22]

[14] OpenAI, "Introducing ChatGPT" (30 November 2022), online: [perma.cc/TJG2-ZM4U].

[15] These are all found in Goudge, *supra* note 3 at 25, 39–40. See a similar catalogue of uses in Beatson, *supra* note 3 at 312; Raso, *supra* note 3 (focusing on risk scores used in the prison security classification context); Re & Solow-Neiderman, *supra* note 3 at 244 (discussing the use of AI for bail and parole determinations).

[16] See e.g. the software at issue in *May v Ferndale Institution*, 2005 SCC 82 at paras 101–20 [*May*] and *Mission Institution v Khela*, 2014 SCC 24 at paras 96–97 [*Khela*] for recommending security classification in prisons; the Supreme Court held in both cases that since weights and values encoded in the software could be disclosed, procedural fairness required this.

[17] Matthew Kosinski, "What is Black Box AI?" (29 October 2024), online: [perma.cc/XHE3-ZDPW]. See also Emily Berman, "A Government of Laws and Not of Machines" (2018) 98:4 BUL Rev 1277 ("the more complex and powerful an algorithm, the more opaque it is likely to be" at 1316).

[18] Kosinski, *supra* note 17.

[19] For a survey of work exploring this distinction, see Alejandro Barredo Arrieta et al, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI" (2020) 58 Information Fusion 82.

[20] Kosinski, *supra* note 17. See also Cary Coglianese & David Lehr, "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era" (2017) 105:5 Geo LJ 1147 ("[t]he results of machine learning analysis are not intuitively explainable and cannot support causal explanations of the kind that underlie the reasons traditionally offered to justify governmental action" at 1167).

[21] Daly, "Artificial Administration", *supra* note 8 ("where [a] decision-maker cannot reach a decision without exercising discretion or judgement, based on the characteristics of the claim being made, it will be difficult to rely on technology alone, because the reasons for the decision arrived at are likely to be opaque" at 102).

[22] Goudge, *supra* note 3 at 29, citing Harry Surden, "Artificial Intelligence and Law: An Overview" (2019) 35:4 Ga St U L Rev 1305 at 1336; Coglianese & Lehr, *supra* note 20 at 1159–60, 1167. On bias, see Tasioulas, *supra* note 10 at 25.

A third broad concern was that AI could not engage in forms of reasoning essential to legal judgment. AI was assumed to lack "higher-order human abilities," including "abstract thinking, conceptual interpretation, and understanding of normative social values."[23] It was also thought to lack "common sense, empathy, or moral reasoning."[24] AI could not effectively apply "discretionary rules that demand appreciation of context," or make "principle-based judgments where there is no 'clear right or wrong answer.'"[25] It could not identify relevant facts based on the statutory framework's "normative policy intents."[26] Nor was AI thought to be suited to make "[d]iscretionary decisions calling for individualised assessment of the circumstances of each applicant."[27]

Assumptions about AI's functional limits can also be found in policy statements on AI in administrative law both prior to 2022 and since.[28] To foreground only one of them, the Treasury Board's *Directive on Automated Decision-Making*, first issued in 2019 and updated in 2023,[29] sets out a four-level classification scheme of possible uses of AI to help decide cases with an impact on individual rights ranging from low to high.[30] The Directive requires at all four levels that an affected person be provided a general description of "the criteria used to evaluate input [or client] data and the operations applied to process it."[31] This assumes that AI could not satisfy concerns about transparency and impartiality other than by disclosures about internal processes, or not by offering cogent reasons for decision. The Directive also mandates that decisions at all but the lowest level "cannot be made without having clearly defined human involvement during the decision-making process" and that the "final decision must be made by a human."[32] This assumes that AI is too risky to be used alone due in part to its lacking necessary reasoning capabilities for judgment noted earlier.[33]

B. NEW CAPABILITIES

The assumption in earlier scholarship and policy that AI involving machine learning entails a measure of opacity or is explainable but not fully interpretable is still current.[34] Yet generative AI involving large language models has brought about a host of new capabilities

[23] Goudge, *supra* note 3 at 29, citing Surden, *supra* note 22 at 1309.

[24] *Ibid*.

[25] Goudge, *supra* note 3 at 29, quoting Surden, *supra* note 22 at 1337.

[26] Goudge, *supra* note 3 at 29.

[27] Daly, "Artificial Administration," *supra* note 8 at 100; Tasioulas, *supra* note 10 at 25 (asserting that AI cannot extrapolate to new situations).

[28] Treasury Board of Canada Secretariat, *Directive on Automated Decision-Making* (Treasury Board Directive), 2019 version (Ottawa: TBS, 2019), online: [perma.cc/E9SC-WLNB]. Another important policy statement on point is the Treasury Board of Canada Secretariat, *Guide on the Use of Generative Artificial Intelligence* (Ottawa: TBS, 2023), online: [perma.cc/K5RD-SDCD] [Treasury Board, *Guide on the Use of AI*].

[29] TBS, *Directive on Automated Decision-Making* (Treasury Board Directive), current version (Ottawa: TBS, 2023), online: [perma.cc/7V8P-2LAG].

[30] *Ibid* at Appendix B.

[31] *Ibid* at Appendix C, accompanying s 6.2.3.

[32] *Ibid* at Appendix C, accompanying s 6.3.13.

[33] The Treasury Board's *Guide on the Use of AI*, *supra* note 28, reflects similar assumptions about AI's functional limitations. Under the heading "Policy considerations and best practices," it states that "generative AI may not be suited for use in administrative decision-making," since the "design and function of generative models can limit institutions' ability to ensure transparency, accountability and fairness."

[34] Kosinski, *supra* note 17; Barredo Arrieta et al, *supra* note 19.

that call other assumptions canvassed here into question: that AI cannot give reasons that justify or render the basis for arriving at a decision intelligible; that without this capability, all AI decisions are potentially biased or flawed; and that AI cannot reason judiciously (that is, be contextually sensitive, engage in normative reasoning, or apply law to individual facts correctly).

Scholars and jurists have begun conducting experiments with language models in both the appellate and administrate law contexts to demonstrate that generative AI is capable of mitigating to a significant degree all three of these concerns.[35] As the results of these experiments demonstrate, a language model provided with sufficient material (more on this below) can produce an opinion setting out cogent and persuasive reasons that justify a decision — that is, making intelligible a rational, if not a technical basis, for it.[36] Over the course of many experiments involving material submitted in actual cases, AI's reasons for decisions were consistently responsive to party submissions; models could make normative judgments and engage in other forms of creative legal thought, such as formulating new legal tests; and models could readily apply legal tests or factors to the circumstances of a case correctly and effectively.[37]

To highlight three of these experimental efforts, the most notable and extensive (to this author's knowledge) is Adam Unikowsky's pioneering experiments in the summer of 2024, which have begun to attract scholarly attention.[38] He uploaded to Anthropic's Claude 3.0 language model briefs from 37 of the United States Supreme Court cases in the current term and asked it to briefly outline a decision in each case in three to four paragraphs.[39] He found that Claude decided 27 of the 37 cases the same way the Court did and that in the remaining 10, he "frequently was more persuaded by Claude's analysis than the Supreme Court's."[40] His write-up of the experiment includes a detailed discussion of six decisions rendered in the previous week.[41] Claude, he writes, "nailed five out of six [of the cases], missing only *Campos-Chaves*, in which it took the dissenters' side of a 5-4 opinion, which is hardly 'wrong.'"[42] Unikowsky documents Claude's ability to formulate different, more elaborate, and more creative legal tests (in relation to those found in the actual decisions) — and the effective application of the tests or factors to the facts in the case at bar.[43] The experiment as a whole presents strong evidence of a language model's capacity to consistently render the

[35] Unikowsky, "Trust", *supra* note 6; Unikowsky, "Trust Part II", *supra* note 6; Diab, *supra* note 6; Daly, "Artificial Intelligence", *supra* note 6.

[36] See e.g. the experiments in Diab, *supra* note 6 using factums submitted in *R v Bykovets*, 2024 SCC 6 [*Bykovets*] and *R v Singer*, SCC 41090 [*Singer*].

[37] Unikowsky, "Trust", *supra* note 6*;* Unikowsky, "Trust Part II", *supra* note 6; Diab, *supra* note 6.

[38] See especially Unikowsky, "Trust", *supra* note 6 and Unikowsky "Trust Part II", *supra* note 6 for Unikowsky's experiments. A brief discussion of them can be found in André Guskow Cardoso et al, "Generative Artificial Intelligence and Legal Decision-making" (2024) 19:11&12 Global Trade & Customs J 710 at 711–12; Diab, *supra* note 6.

[39] Unikowsky, "Trust", *supra* note 6.

[40] Unikowsky, "Trust Part II", *supra* note 6.

[41] *Ibid.* This suggests that Claude was not drawing on training data to arrive at the outcome in these cases; that is to say, Claude's decision was not shaped by the phenomenon of memorization of the actual text in the model itself. Evidence of the technical possibility of this is canvassed in A Feder Cooper & James Grimmelmann, "The Files are in the Computer: On Copyright, Memorization, and Generative AI" (2005) Chicago-Kent L Rev [forthcoming in 2025].

[42] Unikowsky, "Trust Part II", *supra* note 6 referring to *Campos-Chaves v Garland*, 602 US 447 (2024).

[43] Unikowsky, "Trust Part II", *supra* note 6.

outline of a decision in an apex court case that approaches the quality and sophistication of the court's actual decision.[44]

To replicate part of the experiment, I uploaded to OpenAI's GPT 4.5 seven of the factums in *R v. Bykovets*,[45] a case decided by the Supreme Court of Canada in 2023, and nine of the factums in *R v. Singer*,[46] a case heard by the Supreme Court at the time of this writing (mid-2025) but not yet decided.[47] Both experiments confirmed Unikowsky's findings that a language model is effective in carrying out many of the basic reasoning tasks in judgment that scholars were doubtful about. GPT 4.5's draft opinions contained brief but sufficiently nuanced and accurate summaries of the relevant facts and the parties' positions.[48] Each outline then identified the key legal issue or two to be decided, followed by a paragraph outlining a decision on the main issues, with reasons to justify it — reasons closely approximating in quality those of the Supreme Court itself.[49] The reasons in each case also demonstrated an ability to apply a legal test to novel facts.[50]

The key in both Unikowsky's experiment and in my efforts at replicating portions of it is that the language model in each case drew primarily on party materials uploaded to the model. A model can be responsive to party positions, it can be contextually sensitive, and it can apply law correctly to the facts in a given case if provided materials containing these specifics. The crucial advance of generative AI is its ability — once provided the relevant material — to demonstrate sophisticated legal reasoning, or at least the appearance of this (more on the distinction below), to *justify* a decision. What the models do not do is reveal the specifics of the technical process, under the hood, for reaching their output. I set aside for the moment what inferences should be drawn from this fact. The aim here is only to register the point that if provided sufficient material, a model can generate reasons for decision that offer a transparent and intelligible explanation as to why it should be accepted as the correct outcome in both a legal and moral sense — subject to the possibility of review on other grounds, including concerns about bias in underlying processes (more on this below).

One further experiment to note in the administrative law context specifically can be found in recent scholarship by Paul Daly.[51] Daly tested GPT 4's ability to draft reasons for decision in the immigration law context declining a visiting visa.[52] He provided the model with "basic information about Canadian law" and some of the applicant's circumstances, drawing on an actual case brought before the Federal Court for judicial review.[53] Daly's aim was to demonstrate not a language model's ability to decide a case on its own (drawing on party materials), but to assist a decision-maker in outlining draft reasons in accordance with their

[44] This can be gleaned by surveying the results in both Unikowsky, "Trust", *supra* note 6 and Unikowsky, "Trust Part II", *supra* note 6.

[45] *Supra* note 36.

[46] *Supra* note 36.

[47] Outlines from my experiment with *Bykovets* can be found at [perma.cc/Y8BX-DX3B] ["Result page for *Bykovets*"] and with *Singer* at [perma.cc/ZLB8-5JNS] ["Result page for *Singer*"].

[48] See the outlines, *ibid*.

[49] I compare GPT 4.5's outline for *Bykovets* with the Supreme Court's decision in Diab, *supra* note 6.

[50] See the outlines, *supra* note 47.

[51] Daly, "Artificial Intelligence", *supra* note 6 at 16.

[52] *Ibid* at 17–18.

[53] Daly, "Artificial Intelligence", *supra* note 6 at 16; *Haghshenas v Canada (Citizenship & Immigration)*, 2023 FC 464.

instructions.[54] The output that Daly documents is, however, similarly cogent and responsive as was the output in the experiments noted above.[55]

II. *VAVILOV*, REASONABLENESS, AND AI

Vavilov reformed and consolidated the law of judicial review of administrative decisions in Canada.[56] It sought to simplify matters by asserting a presumption that courts should review decisions on the standard of reasonableness, and it defined a set of narrow exceptions when the standard of correctness will apply.[57] The majority justified the presumption of a reasonableness standard as consistent with the broader principle of judicial deference to legislative intent.[58] If a legislature has assigned a decision-making role to a tribunal or a decision-maker without specifying a court's role on review, the majority held that the legislature should be assumed to have intended minimal judicial interference in those decisions, making reasonableness the appropriate presumptive standard.[59] The majority canvassed the application of this standard in detail. In what follows, I consider *Vavilov*'s elucidation of the standard of reasonableness in light of the possibility that a legislature has mandated that high-impact decisions be made by or in reliance upon a language model (in the manner contemplated above). The question here is whether concerns about opacity or the algorithmic or statistical basis of AI decisions hinder them from withstanding a reasonableness review as contemplated in *Vavilov*.

As noted earlier, the majority in *Vavilov* held that in general terms, "a reasonable decision is one that is based on an internally coherent and rational chain of analysis" and "justified in relation to the facts and law that constrain the decision maker."[60] Reasons will justify an outcome if certain conditions are met. These include the requirement that reasons comply with statutory directives and precedent, be responsive to party submissions, and consider the impact of the decision on the affected person.[61] A question not addressed in *Vavilov* is how to apply its key requirement to reasons that AI might provide: namely, that the reasons "justify to the affected party, in a manner that is transparent and intelligible, the basis on which [the decision-maker] arrived at a particular conclusion."[62] When a language model decides a case, what is the *basis* of the decision? Is it the reasons that a model provides in its textual output? Or is it the largely opaque technical process (involving layers of machine learning) that led to the output in question?

The majority in *Vavilov* approaches the question of reasonableness by making a tacit assumption about written reasons. Where reasons are given that purport to justify a decision, they are the basis for the decision. The Supreme Court is silent on the possibility that language models have brought about: providing reasons that make intelligible a rational basis for arriving at a certain outcome, while the actual, technical basis is statistical or algorithmic and

[54] Daly, "Artificial Intelligence", *supra* note 6 (noting the prompt: "Can you generate 'the reasons for refusal based on the applicant's circumstances', based on specifically a belief that the student does not intend to leave Canada after their studies?" at 17).

[55] See the AI decision *ibid* at 17–18.

[56] *Vavilov*, *supra* note 1.

[57] *Ibid* at paras 10, 23, and, summarizing the exceptions, para 17.

[58] *Ibid* at para 24.

[59] *Ibid*.

[60] *Ibid* at para 85.

[61] *Ibid* at paras 108–35.

[62] *Ibid* at para 96. See also *Dunsmuir*, *supra* note 2 at para 47.

to some degree opaque. The question here is whether a court reviewing an AI decision should accept AI's textual reasons as the basis for the decision when assessing whether it was reasonable. Where a court finds that reasons are internally coherent, logical, and responsive to law and fact, should it still be found to be unreasonable due to concerns about underlying opacity? *Vavilov* contains passages supporting either position. I briefly note them before moving to broader theoretical arguments in favour of either view in the final section of this article.

Both the majority and the minority in *Vavilov* confirm that a party challenging a decision bears the burden of showing it to be unreasonable.[63] As the majority notes, this involves persuading a court that a decision contains "serious shortcomings" that prevent it from "exhibit[ing] the requisite degree of justification, intelligibility and transparency."[64] Where a decision-maker has provided written reasons, the majority holds that "those reasons are the means by which the decision maker communicates the rationale for its decision. A principled approach to reasonableness review is one which puts those reasons first."[65] (Note the use of the singular here: "*the* rationale for its decision.") On one reading, then, written reasons are the *rationale* for a decision; internal *processes* for arriving at it (biological on the part of a human, machine learning in AI) may be just that.

But can AI satisfy the substantive requirements for a reasonable decision? The majority provides further direction on what this entails. A reasonable decision must be "internally coherent" and "logical"; as the experiments canvassed above suggest, language models can readily craft coherent and logical reasons explaining a given outcome.[66] A decision must also be justified in terms of law and fact on various specific grounds set out in *Vavilov*.[67] Here too, the experiments noted above show that AI can meet these tests: a model can apply law to facts in a manner that accords reasonably with governing statute and precedent; principles of statutory interpretation; evidence; party submissions; and it can make clear an appreciation of the impact of the decision on the affected individual. The majority also holds that a reasonable decision must avoid reliance on "irrelevant stereotypes" that would give rise to a reasonable apprehension of bias.[68] But what is contemplated here are forms of bias discernable in the reasons themselves (the majority's example being the decision on review in *Baker v. Canada (Citizenship and Immigration)*), and this too is something a model can be prompted to avoid in its reasons, if not in its training data.

The thrust of the majority's pronouncements on what would render a decision unreasonable is a focus on patterns of thought or analysis as set out in written reasons.[69] These

[63] *Ibid* at para 100 in the majority opinion; paras 291 & 305 in the opinion of Abella and Karakatsanis JJ, concurring in the result but dissenting in part.

[64] *Ibid* at para 100.

[65] *Ibid* at para 84. The minority concurs on this point ("[t]he administrative decision is the focal point of the review exercise. Where reasons are provided, they serve as the natural starting point to determine whether the decision-maker acted reasonably" at para 296).

[66] *Ibid* at paras 85, 102.

[67] *Ibid* at paras 105–35.

[68] *Ibid* at para 126, citing *Baker v Canada (Minister of Citizenship and Immigration)*, 1999 CanLII 699 (SCC) [*Baker*].

[69] In *Vavilov*, *supra* note 1 at para 136, the majority considers the review of a decision made without reasons, such as "where a municipality passes a bylaw or a law society renders a decision by holding a vote," but notes that "even in such circumstances, the reasoning process that underlies the decision will not usually be opaque" (*ibid* at para 137). In this case, "a reviewing court must look to the record as a

reasons are presumed to be reasonable from the outset, with the applicant on judicial review bearing the burden of establishing otherwise.[70] A decision presumed to be reasonable implies that it was made without bias. As noted earlier, where humans make administrative decisions, they are presumed to have done so impartially.[71] The focus on written reasons in *Vavilov* — on their coherence and integrity — would suggest that an AI decision could withstand a reasonableness analysis despite a measure of technical opacity in its inner processes if the presumption of impartiality were extended to AI. (But, of course, whether to extend this presumption, whether concerns about opacity can be mitigated sufficiently, are precisely the issues to be sorted out; more on this below.) If, however, the presumption of impartiality were to apply to AI, then on this reading of *Vavilov*, an AI decision could be reasonable.

There are, however, other pronouncements in the majority opinion that give pause to the case for AI. In broad terms, the majority holds that where "a decision maker's rationale for an essential element of the decision is not addressed in the reasons and cannot be inferred from the record, the decision will generally fail to meet the requisite standard of justification, transparency and intelligibility."[72] The majority notes that "an otherwise reasonable outcome also cannot stand if it was reached on an improper basis,"[73] if it involved "an unreasonable chain of analysis,"[74] or if it "contain[s] a fundamental gap."[75] In a close case where a language model decides a case one way over another and either outcome could be reasonable, whether opaque processes underlying the choice would constitute "an essential element of the decision" or an integral but missing part of the "chain of analysis" is a point on which *Vavilov* is silent.[76] But in ways to be explored in the next section, commentators have made arguments along these lines, suggesting that regardless of AI's readable output (its rational justification), the true rationale must be located in its internal processes, since only they can explain precisely how and why it generated the output that it did. On this reading, a language model's decision in a high-impact case cannot meet the standard of reasonableness because while written reasons might provide one form of justification, they would not provide a *fully* "transparent and intelligible" explanation for the outcome at issue.[77]

Before turning to these theoretical concerns about *Vavilov*'s application to AI, it should be noted that there are other ways in which a decision made by or in reliance on AI might violate administrative law principles. That is to say, correctness on certain questions of law or reasonableness under *Vavilov* are not the only concerns here. For example, where a decision made by or in reliance on AI turns on weights and correlations that remain opaque, it might breach duties of procedural fairness. Which of these duties apply in a given situation depends on "statutory, institutional, and social context."[78] But they may include duties to

whole to understand the decision," and "in doing so, the court will often uncover a clear rationale for the decision" (*ibid* at 137). The minority concurs on this point (*ibid* at para 312). I assume that looking to the record as a whole could not apply to resolve concerns about an AI decision in a high-impact case, since reasons are required in those cases and AI would provide a specific rationale.

[70] *Vavilov*, *supra* note 1 at para 100.

[71] *Keita*, *supra* note 13 ("[t]ribunals benefit from a presumption of impartiality and, as such, any allegation of bias must be supported by concrete evidence and cannot be raised lightly" at para 1). See also *Panov*, *supra* note 13 at para 20; *Arthur*, *supra* note 13 at para 8.

[72] *Vavilov*, *supra* note 1 at paras 98, 136.

[73] *Ibid* at para 86.

[74] *Ibid* at para 87.

[75] *Ibid* at para 96.

[76] *Ibid* at para 98.

[77] *Ibid* at para 96.

[78] *Baker*, *supra* note 68 at para 22.

receive notice of the case to meet and disclosure of relevant materials,[79] the right to be heard by an impartial decision-maker, and to receive reasons for a decision.[80] A decision resting on opaque processes that conceal the technical basis for an outcome might conceivably be challenged on one or more of these grounds. Scholars have explored a host of other possible concerns.[81]

The focus here, however, is on whether a language model's underlying non-interpretability and algorithmic basis preclude it from withstanding a challenge on the standard of reasonableness as outlined in *Vavilov*. In a challenge to an AI decision under *Vavilov*, are there compelling reasons to focus primarily on AI's textual reasons rather than its internal processes?

III. CAN AI MEET THE STANDARD OF REASONABLENESS? ARGUMENTS FOR AND AGAINST

If AI does come to be used to a greater degree in administrative law to the point where it decides cases alone in some areas, courts will have to address how the standard of reasonableness in *Vavilov* should apply or be modified for AI. Two questions central to that assessment will be whether to accept a language model's textual reasons as the basis of its decision rather than its underlying process, and whether to adopt a presumption of impartiality to AI under certain conditions. The broader argument in this final portion of the article is that the new capabilities that language models have brought about, along with the possibility of taking measures to mitigate their internal biases, support a conditional presumption of impartiality. This approach is, in turn, more consistent with the thrust of the reasonableness analysis in *Vavilov*, including its many qualifications noted above.

A. ARGUMENTS AGAINST RELYING ONLY ON AI'S TEXTUAL OUTPUT WHEN ASSESSING REASONABLENESS

I foreground two general arguments here. Both are composites of arguments that various scholars have made in relation to AI at different points in its development.

One line of argument concedes that bias in a language model's training data can be mitigated in some measure and that machine learning processes underlying a model can be explained to some degree, if not made fully interpretable.[82] But it would insist that any

[79] See e.g. in both *May, supra* note 16 and *Khela*, *supra* note 16, decisions about an inmate's security classification were found to be unfair for a failure to provide disclosure of the scoring matrices used in the decision process (even though in the latter case, an officer chose to override the score).

[80] See the discussion of these requirements in *Baker*, *supra* note 68 at paras 21–48.

[81] See e.g. Surden, *supra* note 22; Coglianese & Lehr, *supra* note 20; Beatson, *supra* note 3; Goudge, *supra* note 3; Daly, "Artificial Administration", *supra* note 8; Rebecca Williams, "Rethinking Administrative Law for Algorithmic Decision Making" (2022) 42:2 Oxford J Leg Stud 468.

[82] On strategies to assess and mitigate bias in language models see the survey of scholarship in Yufei Guo et al, "Bias in Large Language Models: Origin, Evaluation, and Mitigation" (2024) arXiv:2411.10915v1, online: [perma.cc/J527-694E]; Leif Azzopardi & Yashar Moshfeghi, "PRISM: A Methodology for Auditing Biases in Large Language Models" (2024) arXiv:2410.18906v2, online: [perma.cc/PK7S-48Z4]. On model explainability, see the survey in Haiyan Zhao et al, "Explainability for Large Language Models: A Survey" (2023) arXiv:2309.01029v3, online: [perma.cc/H6K9-3JN3]. See also Srihari Maruthi et al, "Language Model Interpretability – Explainable AI Methods" (2022) 2:2 Austl J Machine Learning Research & Applications, online: [perma.cc/M84D-25GF].

decision a model makes are still being made by processes that can and do conceal "embedded systemic bias."[83] If language models are not fully interpretable and we cannot be certain their decision-making process is free of bias or error, they should not be relied upon where the human stakes are high.[84] To put this in administrative law terms, the possibility of bias in any language model's training data is real or non-speculative rather than hypothetical, and this unavoidably taints any model's decisions with a reasonable apprehension of bias. If any presumption should apply, it should be a presumption of partiality.[85]

A second, more theoretical argument goes beyond a reluctance to extend a presumption of impartiality to language models. It assumes that even if a model's underlying processes were fully interpretable and free of bias, the model could still not render a reasonable decision because its basis would be found in nothing more than "statistical correlations" rather than a normative rationale.[86] Its decision would have a basis but not a rationale per se.[87] AI could thus never provide the "right kind" of justification required in law.[88] The distinction maps onto one traceable to Ludwig Wittgenstein's distinction between following a rule and complying with it.[89] As Ebrahimi Afrouzi has argued, a rationale is a reason offered in response to a normative inquiry, an explanation that might be as simple as "X should be done because a rule, precedent, or other legal value compels it."[90] AI could never decide a case one way rather than another because it is what law or morality compels; it can only make decisions on a correlative basis, discerning patterns among words and syllables (or

[83] Daly, "Artificial Intelligence", *supra* note 6 at 23; see also Berman, *supra* note 17 (describing machine-learning processes as "unavoidably opaque" and "inescapably reflect[ing] the values, biases, and judgment calls of their programmers, often in ways that are invisible on the face of the algorithm" at 1283); Danielle Keats Citron & Frank Pasquale, "The Scored Society: Due Process for Automated Predictions" (2014) 89:1 Wash L Rev 1 ("[b]ecause human beings program predictive algorithms, their biases and values are embedded into the software's instructions" at 4); Goudge, *supra* note 3 at 24; Beatson, *supra* note 3 at 319.

[84] Berman, *supra* note 17 (arguing that while efforts at rendering machine learning processes might render them explainable and "might increase public confidence in the accuracy and fairness of the resulting models, they ultimately do not represent actual transparency and cannot mitigate the rule-of-law-based concerns" at 1318). See also Re & Solow-Neiderman, *supra* note 3 (arguing that AI incomprehensibility based on machine learning gives rise to "legitimacy or fairness problems" at 264); Neli Frost, "The Impoverished Publicness of Algorithmic Decision Making" (2024) 44:4 Oxford J Leg Stud 780 ("[a]lgorithmic decision making is indeed often biased, is typically opaque and unexplainable, and can result in unjust, rights-infringing decisions at least some of the time" at 781).

[85] Treasury Board, *Guide on the Use of AI*, *supra* note 28, in its caution against the use of generative AI to make administrative decisions (see note 33, above) presents an example of policy premised on this argument.

[86] Afrouzi, *supra* note 10 at 370, 377, 387.

[87] *Ibid* (AI decisions are "necessarily deficient in rationale" at 370).

[88] Tasioulas, *supra* note 10 at 33 [emphasis removed]. Tasioulas states that a decision based on machine learning involves a "categorically different process from the essentially normative enterprise of *justifying* a decision in a particular case by reference to the relevant reasons for that decision" (*ibid* [emphasis in original]). See also Afrouzi, *supra* note 10 ("*legally valid* decisions must be correct not only in their holding but also in their rationale, and yet, AI decisions cannot even hope to be correct in their rationale" at 383 [emphasis in original]).

[89] See the discussion of passages in Ludwig Wittgenstein, *Philosophical Investigations*, translated by GEM Anscombe (Prentice Hall 1973) on rule-following versus compliance in Ian R Kerr & Carissima Mathen, "Chief Justice John Roberts is a Robot" (2014) University of Ottawa, Working Paper at 23–27, online (pdf): [perma.cc/6M9R-KX5W].

[90] Paraphrasing Afrouzi, *supra* note 10 at 377, 395 ("all that the rationale of a judgment needs in order to pass it is having some normative material in its mix, be it a rule, a precedent, or some other legal value" at 395); see also Kerr & Mathen, *supra* note 89 at 23 ("the very activity of judging requires following rules" and not merely complying).

"tokens").[91] And as John Tousialis argues, an automated decision that merely complies with rather than follows a rule cannot provide assurance that it was not just "coincidentally congruent with the law," that it was "arrived at *precisely because* it is in accord with the law."[92] The reasons that a language model might provide for a decision in its textual output are merely ex post justifications for it.[93] While they might support the decision, they cannot truly justify it since they were not "causally efficacious" in bringing it about.[94]

Humans might reason in the same way, deciding upon an outcome first and then coming up with reasons to justify it.[95] But, in theory, we can reason to a result by first asking what law or morality require and choosing an outcome on that basis — providing a causal connection.[96] When a human offers an outline of their reasoning process, it may justify the decision because it is the real basis for the outcome, or must be accepted as such in the absence of evidence to the contrary. With language models, this can never be the case.[97] One might argue in response that the distinction here is illusory. At a neurophysiological level, humans also make decisions for reasons that are either opaque or non-normative.[98] The response to this objection is to argue that human decisions are not reducible to processes on which consciousness "supervenes," but are instead made "at the level of conscious human reasoning" — a level to which AI has no equivalent.[99]

The reluctance to accept AI's ex post reasons as the basis of or justification for its decisions gives rise to a host of broader arguments about automated judgment being incompatible with the rule of law or with democratic or cultural ideals.[100] The arguments take

[91] Afrouzi, *supra* note 10 at 375, 377; see also Tasioulas, *supra* note 10 makes a similar argument ("it is highly questionable that … an automated system operates in a way that can be accurately described as 'following legal rules'" at 23). Williams, *supra* note 81 draws a similar distinction ("where [machine learning] is used, it relies on statistical inferences, not reasoning, so when such a system uses a particular factor it is not identifying that factor as relevant to the decision it is making in the way that a human would, but merely recognising that, statistically, that factor often correlates with the relevant outcome" at 487); Berman, *supra* note 17 ("the models and their predictions are based on identified correlations within a data set rather than proven causal relationships" at 1283). Afrouzi, *supra* note 10 at 370, traces the distinction as far back as Mireille Hildebrandt's discussion of the difference between the rule of law and the rule by code-driven law in "Legal and Technological Normativity: More (and Less) Than Twin Sisters" (2008) 12:3 Techné 169.

[92] Tasioulas, *supra* note 10 at 32 [emphasis in original].

[93] Afrouzi, *supra* note 10 at 378–79; Tasioulas, *supra* note 10 at 34.

[94] Tasioulas, *supra* note 10 at 34.

[95] Afrouzi, *supra* note 10 at 386–88.

[96] *Ibid* at 375, 387.

[97] *Ibid* at 383.

[98] Afrouzi, *supra* note 10 at 386–87; Volokh, *supra* note 12 (advancing the view that "[i]f we are honest with ourselves, we often can't really tell with confidence why we reached a particular judgment … We have reactions because of the real neural nets in our brains, and then we can offer explanations that we hope persuade" at 1164–65). See also Berman, *supra* note 17 (exploring "contexts where human decision-making is itself opaque" at 1319).

[99] Afrouzi, *supra* note 10 at 387.

[100] See e.g. Berman, *supra* note 17 (exploring "whether reliance on the output of machine-learning models—even if highly accurate—is in tension with the goal to maintain 'a government of laws' and not of machines" at 1282); Re & Solow-Neiderman, *supra* note 3 ("[b]y offering efficiency and at least an appearance of impartiality, AI adjudication will foster a turn [away from 'equitable justice'] toward 'codified justice,' that is, a paradigm of adjudication that favors standardization above discretion" at 246). See also Frost, *supra* note 84 (arguing "the increasing deployment of [machine learning] technologies compromises the publicness of administrative decision making and decisions to generate an *impoverished* publicness and thereby destabilise this site's democratic potential" at 782 [emphasis in original]).

two forms. Various authors contend that decisions are (or should be rejected as) incompatible with values, models, or concepts of what makes for legitimate legal judgment in a self-governing social or political order.[101] The other form of the argument is focused on belief and opinion. Using AI to make decisions in serious cases is unlikely ever to be accepted as legitimate or desirable. This is due to the widely held intuition that machines are unsuitable for legal judgment due to the rote, vacuous, or merciless way in which they make determinations.[102] I note these arguments here without exploring them given that many take as their point of departure a distinction between correlative and normative bases for judgment.

B. ARGUMENTS FOR RELYING ONLY ON AI'S TEXTUAL OUTPUT WHEN ASSESSING REASONABLENESS

Two general lines of argument can be offered in response. One argues for treating a language model's output as presumptively impartial; the other for treating a model's textual reasons as the basis for a potentially reasonable decision as contemplated in *Vavilov*.

To recall the first of the arguments canvassed above, the concern was that since models are known to carry a measure of bias in their training data and their outcome cannot be fully interpretable, they cannot overcome a reasonable apprehension of bias. One response might begin with the fact that language models come in many shapes and sizes. Nothing precludes a model from being developed and run by a tribunal in-house, where officials can control and oversee fine-tuning, auditing, and testing of the model.[103] Through fine-tuning, the use of system prompts, or techniques such as reinforcement learning, a model's likelihood of generating biased, erroneous, or harmful content in its output can be significantly reduced — in ways that can be tested and through processes that can be understood.[104] These steps could, in theory, quell concerns about bias and explainability or render them mostly hypothetical.[105] One might object to this line of argument by suggesting that it is tantamount to what is called "computational reliabilism," or the view that an automated decision is justified if its processes

[101] See e.g. Afrouzi, *supra* note 10 at 383; Daly, "Artificial Administration", *supra* note 8 at 102–04, discussing models of administrative justice drawing on Jerry L Mashaw, *Bureaucratic Justice: Managing Social Security Disability Claims* (New Haven: Yale University Press, 1983); Tasioulas, *supra* note 10 (arguing that "AI-based decision-making might inhibit the full realisation of the rule of law" on various grounds at 29).

[102] Daniel Kahneman, Olivier Sibony & Cass R Sunstein, *Noise: A Flaw in Human Judgment* (New York: Little, Brown Spark, 2021) at 340 (arguing that lacking a capacity for mercy, AI is not capable of judging with dignity). See also Re & Solow-Neiderman, *supra* note 3 ("[b]ecause of its discretionary, contextual, and dynamic nature, equitable justice may seem flatly incompatible with automated algorithmic processes. For instance, can a prearranged decision procedure really incorporate an idea like mercy or develop fact-sensitive balancing of mitigation factors in a criminal case?" at 253).

[103] The Treasury Board's *Guide on the Use of AI*, *supra* note 28, does not mandate that any model used be run in-house, but this would be consistent with its "Recommended approach," which includes "making sure generated content is accurate, legal, ethical, and compliant with the terms of use … [by] establish[ing] monitoring and oversight mechanisms" and "ensur[ing] that the infrastructure and tools are appropriate for the security classification of the information and that privacy and personal information are protected."

[104] See Anthropic's description of its use of system prompts — implicit directions that a model follows in each output — tailored to a set of humanistic principles contained in what it calls "Constitutional AI", including e.g. "DO NOT choose responses that exhibit toxicity, racism, sexism or any other form of physical or social harm": Anthropic, "Claude's Constitution" (9 May 2023), online: [perma.cc/DG88-3ZYX]; for more technical detail see Yuntao Bai et al, "Constitutional AI: Harmlessness from AI Feedback" (2022) arXiv:2212.08073, online: [perma.cc/JA42-6LZE].

[105] See the sources in above note 82.

and output are confirmed to be reliable.[106] It is merely an attempt to offer reliability as a substitute for normative justification. But this would be incorrect in the case of a model that gave reasons, since a model's decision would be assessed in light of those reasons — *along with* evidence of measures taken to ensure reliability or impartiality.[107]

A situation might be envisaged in which a statute and regulations mandate measures to be taken to ensure a degree of bias-mitigation and model explainability, as conditions for the operation of a presumption of impartiality. This might work in a similar way to which *Criminal Code* provisions provide for the use of breathalyzers and devices to measure drug concentrations in breath or blood samples.[108] Presumptions operate if certain conditions as to the maintenance and operation of these devices are met.[109] *Criminal Code* provisions differ in the sense that readings from these instruments are admitted as evidence if conditions are met, rather than being used to make a decision about guilt. Yet they are similar to what is contemplated here in the sense that ever-present concerns about reliability do not preclude a reliance on the instruments where their output, in practical terms, can be close to conclusive of guilt.

The second argument above was that a language model, even if reasonably and reliably free of bias or error, should not be used to decide cases where the stakes are high because their decisions can only be based on statistical correlations rather than on what law or morality compel. They cannot therefore be legitimate, because we assume (in a socio-cultural sense) that valid legal decisions are those made on a normative or empathetic or rational basis that is made clear through reasons. Eugene Volokh, among others, anticipated this argument in 2019, when the prospect of AI acquiring the ability to formulate detailed reasons for decision seemed, to most, a distant possibility.[110] Volokh postulated that at some point, AI would be capable of deciding a case with reasons for decision good enough to convince a panel of human judges in a form of the Turing test.[111] It may do so on the basis of complying with, rather than following a rule, and it may only amount to an ex post justification.[112] But this should not matter.[113] The only concern should be that "an *AI judge's opinions* persuade observers who expect opinions to be consistent with the legal rules. Rule-following is as rule-following does."[114]

[106] T David Tan, "The Legality of Black Boxes in Administrative Law," in Yee-Fee Ng & Matthew Groves eds, *Automation in Public Governance: Theory, Practice and Problems* (Hart, Oxford, 2025) at 5. See also Maya Krishnan, "Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning" (2019) 33:3 Philosophy & Tech 487; Juan M Durán & Nico Formanek, "Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism" (2018) 28:4 Minds & Machines 645.

[107] Tan, *supra* note 106 at 2, 22 concedes this possibility, focusing his critique of computational reliabilism on cases "the required evidence for a legal decision is the reliability of the black box rather than an explanation of the reasons for the decision" (*ibid* at 8) — that it to say, where we have only a black box to rely upon.

[108] RSC 1985, c C-46, s 320.28 [*Criminal Code*].

[109] *Ibid,* ss 320.14(1), 320.31(1), 320.31(4).

[110] Volokh, *supra* note 12; see also Susskind, *supra* note 12, ch 27.

[111] Volokh, *supra* note 12 at 1137–38.

[112] *Ibid* at 1161, 1164.

[113] *Ibid* at 1161.

[114] *Ibid* [emphasis in original].

Volokh's argument for this pragmatic interpretation of judgment rests on a further claim he advances. The only means by which we can determine whether a decision is made on a normative basis — whether it *should* in fact compel an outcome one way or another — is persuasion.[115] More broadly, where either a human or AI renders a decision, the primary basis (though not the only one) on which to assess its legitimacy or acceptability is how persuasive we find its reasons.[116] The same logic applies to compassion or mercy. An AI judge may not possess or feel compassion, but it can produce reasons that express compassion.[117] And here too "what matters is the result, not the process."[118] If AI can generate cogent and persuasive reasons for decision, they should be accepted as legitimate on that basis, in the absence of evidence of bias or error in underlying processes.

Afrouzi's response to Volokh is to assert that if ex post justifications are the best that AI can do, it fails to meet the "legal ideal" we share that judgment should be based on normative reasons.[119] But as Volokh and Richard Susskind have both argued, beliefs that shape our ideas about legitimacy in judgment are culturally or historically contingent.[120] Volokh conceded that judgment may be something people remain reluctant to delegate to AI, but he cautioned being too quick to be certain about this "just because many people's first reaction to the [idea of AI judges] may be shock or disbelief."[121] Similarly, for Susskind, "it is probable that our grandchildren will have different views from ours" on "the computer judge" since "they will live in an age when it will be commonplace for machines to be unarguably superior to humans in many walks of life."[122] Over time, just as using chatbots in therapy has become more widely embraced, the AI judge may soon become more acceptable.[123] Arguments about the fundamental incompatibility of AI judging in administrative law due to a mismatch of values or models (bureaucratic, humanistic, and so on) are, on this view, subject to change.

Stepping back from this debate, we can surmise that the two schools of thought on AI judging are at an impasse. Differences of opinion on what forms the basis of an AI decision are not likely to be resolved in the short term. For some, no matter how well-reasoned or humane an AI decision might appear to be, its technical foundation in underlying statistical or algorithmic processes can never overcome concerns about limited interpretability or the lack of a basis in normative judgment. The argument here for favouring one approach over the other rests on the belief — subject to doubt — that concerns about language model bias and opacity could be mitigated to a reasonable degree to warrant a presumption of impartiality as to its output. And if so, the claim here is that treating a model's reasons as the "basis" of its decision for the purpose of a reasonableness review would conform more closely with the thrust of the Supreme Court's holding in *Vavilov*. As the majority held, where reasons are provided, "[a] principled approach to reasonableness review is one which puts those reasons

[115] *Ibid* ("[i]n ordinary litigation, the winning side is the side that persuades the judge, even if there is no logical way to prove that the winning answer is correct" at 1153).

[116] *Ibid* at 1152.

[117] *Ibid* at 1167.

[118] *Ibid* at 1189. See also Susskind, *supra* note 12 (advocating a form of "outcome-thinking," which "inclines us to consider whether machines can deliver decisions at the standard of human judges or higher, not by replicating the way that judges think and reason but by using their own distinctive capabilities (brute processing power, vast amount of data, remarkable algorithms)" at 280).

[119] Afrouzi, *supra* note 10 at 379.

[120] Volokh, *supra* note 12 at 1171; Susskind, *supra* note 12 at 292.

[121] Volokh, *supra* note 12 at 1171.

[122] Susskind, *supra* note 12 at 292.

[123] Anisha Sircar, "Could A Bot Be Your New Therapist? How AI Has Transformed Mental Healthcare", *Forbes* (28 October 2024), online: [perma.cc/WE2P-QH8Q].

first."[124] Despite the majority's qualifications about "fundamental gap[s]" precluding a finding of reasonableness, a model that gives cogent reasons, responsive to law and fact, in the absence of evidence of bias or error in its inner workings, should be capable of being found reasonable.[125]

IV. CONCLUSION

This article has sought to examine a question not on the Supreme Court's horizon at the time it decided *Vavilov*, but one that is gaining relevance with recent developments in AI. Could an administrative decision made by AI alone withstand a challenge for reasonableness under *Vavilov*? The issue on which the case is silent is how to construe new capabilities that language models have brought about to provide cogent, responsive, and legally correct reasons for decisions. A skeptical view sees problems with bias and opacity as insurmountable and deeper issues with AI's inability to decide cases on a normative basis. An optimistic view of AI's utility sees pathways to attaining a sufficient degree of reliability to warrant a presumption of impartiality and good reason to focus on AI's textual justification over the process by which it was produced. This focus on the quality of written reasons is, in turn, more consistent with the thrust of *Vavilov*, which conceives of reasonableness in terms of a performance of justification — a performance at which AI is becoming increasingly adept.

[124] *Vavilov*, *supra* note 1 at para 84.

[125] *Ibid* at para 96.