Dalhousie Law Journal

Volume 49 | Issue 1 Article 8

7-2026

A Language Model's Capability to Make Reasoned Decisions in Administrative Law

Robert Diab
Thompson Rivers University Faculty of Law

Robert Diab*

A Language Model's Capability to Make Reasoned Decisions in Administrative Law

This paper calls into question a common set of assumptions about the use of artificial intelligence in administrative law in Canada. While some assume that AI (AI) may be useful for assisting a decision-maker where core rights are affected, the prevailing view is that it cannot be relied upon to make final decisions, due mainly to functional limitations: AI cannot give reasons and its outcomes are fraught with bias and opacity that cannot be overcome. The prevailing view, however, took shape in response to earlier, more limited forms of AI. Large language models can be used in a way that mitigates earlier concerns to a significant degree—but both scholarship and policy have yet to absorb this. Experiments canvassed in this paper involving appellate court decisions show that when prompted to draw primarily on party materials, AI can produce a reasoned decision comparable in quality to that of a human, calling for a reassessment of AI's possible role in administrative law and a shift of concern from underlying technical processes to the quality and cogency of reasons that AI can provide.

Le présent article remet en question un ensemble d'hypothèses courantes concernant l'utilisation de l'intelligence artificielle (IA) en droit administratif au Canada. Même si certains estiment que l'IA peut être utile pour aider un décideur lorsque des droits fondamentaux sont en jeu, l'opinion dominante est qu'on ne peut s'y fier pour prendre des décisions définitives, principalement à cause de ses limites fonctionnelles : l'IA ne peut pas fournir de motifs et ses résultats sont truffés de préjugés et d'opacité impossibles à surmonter. Toutefois, l'opinion dominante s'est forgée en réponse à des formes antérieures et plus limitées d'IA. De grands modèles de langage peuvent être utilisés de manière à atténuer considérablement les préoccupations antérieures, mais ni la doctrine ni les politiques ne tiennent encore compte de cela. Les expériences présentées dans cet article, portant sur des décisions de cours d'appel, montrent que lorsqu'on lui demande de s'appuyer principalement sur les documents des parties, l'IA peut produire une décision motivée d'une qualité comparable à celle d'un être humain, ce qui nécessite de réévaluer le rôle possible de l'IA en droit administratif et de se préoccuper non plus des processus techniques sous-jacents, mais de la qualité et de la pertinence des motifs que l'IA peut fournir.

* Professor, Faculty of Law, Thompson Rivers University. I wish to thank Professor Constance MacIntosh and the anonymous reviewers of this paper, along with Lynda Corkum and the student editors of this journal.

Introduction

Scholars and policy-makers in Canada have grappled with the role of artificial intelligence (AI) in administrative law for decades.[1] Their work forms a subset of a larger body of literature on the role of AI in legal judgment more broadly.[2] In both cases, a strong consensus formed prior to 2022 still persists: that given AI's functional limitations, in cases involving core rights—liberty, equality—AI may be useful in helping to draft reasons for a decision, but should not be relied upon to make a decision or decide a case on its own.[3]

Yet the consensus about AI lacking the capacity for judgment took shape in response to earlier, more limited forms of AI. Much of the scholarship on point deals with AI that predicts scores or probabilities of

1. This material is canvased in Part I below.
2. See the comprehensive overview in Ignacio N Cofone, "AI and Judicial Decision-Making" in Florian Martin-Bariteau & Teresa Scassa, eds, *Artificial Intelligence and the Law in Canada* (Toronto: LexisNexis Canada, 2021) 332.
3. This consensus is canvassed in scholarship and policy instruments in Part I below.

an outcome based on a discrete set of variables: for example, a credit score or the risk of reoffending.[4] Scholars of administrative law have thus argued that in high-impact cases, AI decisions could not satisfy principles of procedural fairness or reasonableness—or administrative justice more broadly[5]—based on core limitations in the way that AI functions.[6] AI could not provide reasons for decision; could not be sensitive to social context; could not engage in moral or normative reasoning; and could not identify relevant legal factors when exercising discretion.[7] Recent scholarship and policy statements on the use of generative AI continue to reflect these concerns, including the belief that a degree of bias and opacity underlying the output of AI cannot be overcome.[8]

Generative AI, however, has brought about a sea-change in capabilities. Large language models can now be used in a way that mitigates to a significant degree many if not most of the earlier concerns about AI's capacity for judgment—with implications that policy and scholarship on AI in administrative law have yet to fully absorb.[9] The paper explores these new capabilities by canvassing experiments in which language models

4. Cofone, *supra* note 2 at 333. Examples in administrative law include Amy Goudge, "Administrative Law, Artificial Intelligence, and Procedural Rights" (2021) 42 Windsor Rev Leg Soc Issues 17; Jennifer Raso, "AI and Administrative Law" in Martin-Bariteau & Scassa, *supra* note 2, 182; Jesse Beatson, "AI-Supported Adjudicators: Should Artificial Intelligence Have a Role in Tribunal Adjudication?" (2018) 31 Can J Admin L & Prac 307.

5. Michael Adler, "Understanding and Analyzing Administrative Justice" [Adler, "Understanding Administrative Justice"] in Michael Adler, ed., *Administrative Justice in Context* (Hart, Oxford, 2010) 129, defining this at 129 as "the justice inherent in decision making." See the discussion of this concept in relation to AI in Paul Daly, "Artificial Administration: Administrative Law, Administrative Justice and Accountability in the Age of Machines" (2023) 30:2 Australian Journal of Administrative Law 95 [Daly, "Artificial Administration"].

6. Goudge, *supra* note 4 at 28-29; Beatson, *supra* note 4 at 323; Daly, "Artificial Administration," *supra* note 5 at 10.

7. See e.g. Goudge, *supra* note 4 at 29; Harry Surden, "Artificial Intelligence and Law: An Overview" (2019) 35:4 Ga St U L Rev 1305 at 1309; Cary Coglianese & David Lehr, "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era" (2017) 105:5 Geo LJ 1147 at 1172-1173.

8. Goudge, *supra* note 4 at 30; Surden, *supra* note 7 at 1336; Raso, *supra* note 4 at 194-196; Amy Salyzyn, "Canadian Courts and Generative AI: Broadening Our Gaze to Potential Corrosive Risks," *Slaw* (3 December 2024), online: <slaw.ca/2024/12/03/canadian-courts-and-generative-ai-broadening-our-gaze-to-potential-corrosive-risks/> [perma.cc/P3U6-Y9V6]. See also the discussion of policy instruments in Part I below.

9. A notable exception can be found in Paul Daly, "Artificial Intelligence and Administrative Tribunals" in Yee-Fui Ng & Matthew Groves, eds, *Automation in Public Governance: Theory, Practice and Problems* (Oxford, UK: Hart Publishing, 2025) 95, [Daly "Artificial Intelligence"], discussing possible uses of generative AI to produce draft reasons when prompted to decide a case one way (but not AI's ability to decide a case alone by drawing on party materials). See also Noel Semple, "Could Artificial Intelligence in Decision-Writing Improve Access to Justice?," *Slaw* (10 June 2025), online: <slaw.ca/2025/06/10/could-artificial-intelligence-in-decision-writing-improve-access-to-justice/> [perma.cc/M5XB-X5SZ].

are used to produce draft judgments in apex courts and other cases.[10] The experiments show how models prompted to draw primarily on materials submitted by the parties can produce a reasoned decision comparable in quality to that of a human judge. The reasons demonstrate AI's ability to be sensitive to social context, to make normative judgments, and to decide a case based on relevant legal factors alone. This vastly more expansive and effective capability compels a reassessment of the role that AI can and should play in administrative law and justice. It points to a possible future in which AI plays a greater role in decision-making on the premise that reviewability should turn not on the explainability of underlying processes but on the nature and quality of the reasons provided.

The paper proceeds in three parts. To lend context, Part I briefly canvasses earlier scholarship casting doubts on AI's capability to satisfy administrative principles due to functional limits of earlier forms of AI, and the persistence of these doubts in key policy instruments and commentary to the present. Part II explores new capabilities of generative AI to mitigate the many concerns raised in response to the earlier forms of AI by canvassing experiments described above. It then considers the relevance of these experiments in the administrative law context. Part III outlines arguments for and against using AI to make or draft high-impact decisions.

I. *The earlier consensus about AI's functional limitations*

Concerns about AI in administrative law relate to both the nature of the technology and to broader legal principles on which decisions tend to be challenged. It may be helpful, as a preface to the overview in this section, to articulate basic assumptions on the part of scholars and policy-makers on each of these two fronts.

Some automated decisions are based on a static algorithm or body of code that follows a clear set of rules. These can be disclosed, readily understood, and examined for bias.[11] More elaborate forms of AI, such as

10. Adam Unikowsky, "In AI we trust" (June 8, 2024), online (newsletter): <adamunikowsky.substack.com/p/in-ai-we-trust> [perma.cc/7FPJ-6SNU] [Unikowsky, "Trust"]; Adam Unikowsky, "In AI we trust, part II" (June 16, 2024), online (newsletter): <adamunikowsky.substack.com/p/in-ai-we-trust-part-ii> [perma.cc/6EME-RDBF] [Unikowsky, "Trust Part II"]; Robert Diab, "The Evolving Role of AI in Legal Judgement" (2026) 18:1 L Innovation & Tech 265 [Diab, "Evolving Role of AI"]; Kimo Gandall, Jack Kieffaber, & Kenny McLaren, "We Built Judge.AI And You Should Buy It" (28 January 2025), online: <ssrn.com/abstract=5115184> [perma.cc/6KL3-K3MS]; and Eric A Posner & Shivam Saran, "Judge AI: Assessing Large Language Models in Judicial Decision-Making" (2025) University of Chicago Coase-Sandor Institute for Law & Economics, Research Paper No 25-03, online: <ssrn.com/abstract=5098708> [perma.cc/5TJW-VDMN].

11. The Supreme Court of Canada in *May v Ferndale Institution*, 2005 SCC 82 [*May*], and *Mission Institution v Khela*, 2014 SCC 24 [*Khela*], assumes this possibility in relation to the software at issue

large language models, involve machine learning processes comprised of "multilayered neural networks"—sometimes thousands of layers deep—that render the models "inherently opaque."[12] Developers may know in a broad sense "how data moves through each layer of the network," but not "the specifics."[13] Models are thus explainable but not interpretable.[14] This makes it difficult if not impossible to discern precisely how a model has made a decision in a given case, i.e. "the factors it weighs and the correlations it draws."[15]

The opacity in machine learning can become an issue in administrative law in relation to the two general grounds on which decisions tend to be challenged: reasonableness and procedural fairness. As the Supreme Court held in *Vavilov*, a decision will be unreasonable where it fails to "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."[16] This involves showing that a decision is "based on an internally coherent and rational chain of analysis and that [the decision] is justified in relation to the facts and law that constrain the decision maker."[17] Can a decision that AI renders that is supported by cogent reasons—but rests on machine learning processes that are to some degree opaque—be said to make intelligible the "basis" of the decision?[18] *Vavilov* also holds that a decision can be unreasonable where it does not comply with statutory directives and precedent, is not responsive to party submissions, or fails to adequately consider the impact of the decision on the affected person.[19] Is AI capable of providing reasons that are attuned and responsive in all these ways?

An administrative decision might also breach duties of procedural fairness.[20] The content of this duty in a given situation depends on

---

in those cases.

12. Matthew Kosinski, "What Is Black Box AI and How Does It Work?" (29 October 2024), online: <ibm.com/think/topics/black-box-ai> [perma.cc/562Z-SSF6].

13. *Ibid*.

14. 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.

15. Kosinski, *supra* note 12. See also Coglianese & Lehr, *supra* note 7 at 1167, noting "[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."

16. *Canada (Minister of Citizenship and Immigration) v Vavilov*, 2019 SCC 65 at para 96 [*Vavilov*].

17. *Ibid* at para 85.

18. See Robert Diab, "*Vavilov* and Generative AI" (2025) 63:1 Alberta L Rev [Diab, "*Vavilov* and Generative AI"], exploring the holding in *Vavilov*, *supra* note 16 on reasonableness in relation to the question of whether a language model's textual reasons should be construed to form the "basis" of the decision for review purposes, or whether its basis must be located in underlying technical processes.

19. *Vavilov*, *supra* note 16 at paras 108-135.

20. *Baker v Canada (Minister of Citizenship and Immigration)*, 1999 CanLII 699 (SCC) [*Baker*].

"statutory, institutional, and social context."[21] However, it tends to include duties to receive notice of the case to meet and disclosure of relevant materials, the right to be heard by an impartial decision-maker, and to receive reasons for a decision.[22] A language model might reproduce bias contained in its training data or in the design and operation of the model itself, and both can remain undetected due to the model's opacity.[23] Where a decision made in reliance on AI turns on weights and correlations that remain opaque, it might also breach the duty to provide notice of the case to meet.[24] More broadly, the phenomenon of automation bias—the tendency to become uncritically trusting of automated output—could form the basis of a claim that a decision-maker acting in reliance on AI has unlawfully or unreasonably subdelegated the decision.[25] Scholars have explored a host of other possible concerns on grounds of fairness.[26]

Part II of this paper will explore how new capabilities of large language models—new ways of using AI to generate a decision with reasons—provide a basis for mitigating many of these concerns. Before turning to that discussion, in what follows, I briefly trace a consensus in scholarship and policy as to how AI's functional limitations precluded its use for high-impact administrative decisions on the legal grounds canvased above. The intent here is not to present an exhaustive survey, but to identify commonly perceived limitations of AI and to highlight their basis in an assessment of the technology at an earlier stage of its development.

1. *Scholarly concerns*

Much of the scholarship about AI in administrative judgement—as with the literature on AI in legal judgment more broadly—predates the public release of ChatGPT in 2022.[27] A common feature across both of these earlier bodies of literature is a focus on forms of AI that are limited to generating predictions or scores based on a series of questions or similarities among a set of variables. Common examples in the scholarship on AI in administrative law include software for predicting recidivism (by

21. *Ibid* at para 22.
22. See the discussion of these requirements in *Baker*, *supra* note 20 at paras 21-48.
23. Kosinski, *supra* note 11; Arrieta et al, *supra* note 14 at 83; Goudge, *supra* note 4 at 30-31.
24. For example, in both *May*, *supra* note 11, and *Khela*, *supra* note 11, 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).
25. Daly, "Artificial Administration," *supra* note 5 at 6.
26. See e.g. Surden, *supra* note 7; Coglianese & Lehr, *supra* note 7; Beatson, *supra* note 4; Goudge, *supra* note 4; Daly, "Artificial Administration," *supra* note 5; Rebecca Williams, "Rethinking Administrative Law for Algorithmic Decision Making" (2022) 42 Oxford J Leg Stud 468.
27. For an overview of the pre-2022 literature on AI in legal judgment, see Cofone, *supra* note 2.

parole boards), deciding immigration eligibility, detecting tax evasion or welfare fraud, or allocating social benefits.[28] Authors commonly draw the distinction, noted above, between systems involving "closed-rule" algorithms and machine learning—seeing the latter as a greater concern, given the issues arising from opacity.[29] Yet the AI involving machine learning here is conceived in limited terms, as the foundation of tools for generating scores or predictions. Scholars highlight three general problems with using AI in this more limited form to make administrative decisions.

The first is that AI involving machine learning conceals the technical basis or process by which it arrives at a decision, making it impossible to ascertain whether AI's process was biased or otherwise improper. As one scholar writes: "[w]ithout 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."[30] A second problem is that AI could not itself provide reasons, leaving the basis for a decision unclear. As another scholar writes: "[d]ecision-making by rote, powered by machines, will struggle to provide reasons that engage with the specifics of an individual case, making them vulnerable to high-intensity rationality review."[31] Or again, "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."[32]

Earlier scholarship is still current on the point that machine learning involves opacity or limited explainability.[33] But in ways to be canvassed below, language models call into question the assumption that AI cannot communicate reasons that justify an outcome as unbiased or that they cannot account for all relevant factors in arriving at a decision. With generative AI, the problem of opacity or a lack of interpretability will come to be separable from the ability to provide reasons. In contrast to more limited forms of predictive, score-generating AI, language models can

28. These are all found in Goudge, *supra* note 4 at 24-25, 39-40. See a similar catalogue of uses in Beatson, *supra* note 4 at 312. Raso, *supra* note 4, focuses on risks scores used in the prison security classification context.
29. Goudge, *supra* note 4 at 24; Daly, "Artificial Administration," *supra* note 5 at 3; Beatson, *supra* note 4 at 308.
30. Goudge, *supra* note 4 at 29, citing Surden, *supra* note 7at 1336 and Coglianese & Lehr, *supra* note 7 at 1159-1160, 1167.
31. Daly, "Artificial Administration," *supra* note 5 at 10.
32. *Ibid* at 11.
33. Kosinski, *supra* note 12; Arrieta et al, *supra* note 14.

provide cogent and persuasive reasons for judgment that apply relevant factors and appear alive to concerns about bias—yet they do not reveal the specifics of the *technical* process for doing any of this.[34]

A third common concern with AI's functional limitations in the earlier scholarship was that AI could not engage in forms of reason essential to legal judgment. AI was assumed to lack "high-order human abilities," including "abstract thinking, conceptual interpretation, and understanding of normative social values."[35] AI lacked "common sense, empathy, or moral reasoning."[36] It 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.'"[37] It could not identify relevant facts based on the statutory framework's "normative policy intents."[38] AI was not suited to making "[d]iscretionary decisions calling for individualised assessment of the circumstances of each applicant."[39] Nor was AI's function in this regard likely to improve: "[b]y design, algorithmic devices (whether code- or data-driven) may never be responsive enough to an individual's circumstances to appear 'open minded.'"[40] And confronting novelty was thought to be well beyond the scope of automation.[41]

These observations were accurate given the limited capabilities of the AI tools in view at the time. But here too, in ways to be explored, generative AI can offer reasons for decisions that *appear* to demonstrate these abilities—to be sensitive to context, to make normative judgments, and to apply law to novel facts.

However, in scholarship since 2022, doubts have persisted about AI's suitability to decide cases in high-impact scenarios based on assumptions about its functional limitations.[42] While some scholars are alive to

34. The experiments canvassed in Part II below are intended to demonstrate this.
35. Goudge, *supra* note 4 at 29, citing Surden, *supra* note 7 at 1309.
36. *Ibid*.
37. Goudge, *supra* note 4 at 29, citing Surden, *supra* note 7 at 1337.
38. Goudge, *supra* note 4 at 29.
39. Daly, "Artificial Administration," *supra* note 5 at 7.
40. Raso, *supra* note 4 at 194.
41. Beatson, *supra* note 4 at 323, noting in passing "[d]espite AI's analytical advantages, human adjudicators will likely be much better in the scenario of a novel case. This is where human creativity and judgment becomes a significant asset." See also Coglianese & Lehr, *supra* note 7 at 1173, arguing similarly: "Machine-learning algorithms cannot directly make the choices about these different aspects of a rule's content not only because some of these choices are normative ones, but also because learning algorithms are merely predictive and thus unable to overlay causal interpretations on the relationship between possible regulations and estimated effects."
42. See e.g. Daly, "Artificial Intelligence," *supra* note 9 at 22-23 noting the possibility of "embedded systemic bias." See also David Tan, "The Legality of Black Boxes in Administrative Law," in Ng & Groves, *supra* note 9, arguing that opacity precludes full reliance on AI for making administrative decisions, but conceding at the end of the chapter "[i]f we have language models that provide reasons as well as humans, then many of the objections presented here disappear."

possibilities of using language models to help draft decisions based on human instruction, using them to adjudicate a case alone based on party submissions (and other material) remains underexplored.[43] Suffice it to note here, however, that some of the concerns about AI in the earlier scholarship remain current while some do not. AI based on machine learning, including generative AI, is still opaque (explainable to an extent but not precisely interpretable); yet, in ways to be explored, language models can be used in ways that mitigate concerns about bias and opacity, calling earlier assumptions into question.

2. *Policy instruments*

Assumptions about the limited use to be made of AI in administrative judgment resting on earlier conceptions of AI's capabilities can also be discerned in key policy documents before and after 2022. I briefly note two of them to foreground the assumptions at issue.

A guideline that continues to shape policy across the administrative landscape in Canada is the Treasury Board's *Directive on Automated Decision-Making*.[44] First issued in 2019 and updated in 2025, the *Directive* contemplates the possible use of AI to make a decision "in whole or in part," so long as notice of this is provided,[45] affected persons receive a "meaningful explanation…of how and why the decision was made,"[46] and it accords with a four-part classification scheme.[47] An appendix sets out "Impact Assessment Levels" ranging from level 1, in which a decision will have "little or no impact on rights of individuals" including dignity and privacy, with impacts that are "reversible and brief," to level 4, in which a decision would likely have "very high impacts" on such rights that are "irreversible and perpetual."[48]

Assumptions tied to an earlier phase of AI are implicit in two key requirements in the *Directive*. One pertains to the requirement of humans-in-the-loop for decisions at levels 3 and 4. In these cases, "[d]ecisions

43. Scholars considering the merits of using language models to assist with drafting reasons include Paul Daly in "Artificial Intelligence," *supra* note 9 at 16, discussing his experiment prompting GPT-4 with "basic information about Canadian law" and some of the applicant's circumstances in *Haghshenas v Canada (Minister of Citizenship and Immigration),* 2023 FC 464, to demonstrate the model's capability of outlining a reasoned decision. See also Semple, *supra* note 9, advocating the use of generative AI to assist in drafting reasons for a decision already arrived at.

44. Treasury Board of Canada Secretariat, *Directive on Automated Decision-Making* (Ottawa: TBS, first published 5 February 2019, last modified 24 June 2025), online: <tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592> [perma.cc/VWX5-PE6F] [Treasury Board, *Directive*].

45. *Ibid*, s 6.2.1

46. *Ibid*, s 6.2.3.

47. *Ibid*, s 6.2.1.

48. *Ibid*, Appendix B.

cannot be made without having clearly defined human involvement during the decision-making process," and "[t]he final decision must be made by a human."[49] This reflects an implicit assumption that AI is too risky to be used without human involvement at these levels due in part to its lacking necessary reasoning capabilities for judgment. The *Directive* also 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."[50] This assumes both the possibility of obtaining a measure of explainability of internal processes that would satisfy concerns about transparency and impartiality—and that AI could not satisfy those concerns in its reasons for decision (a point explored further in Part II below).

A more recent and notable set of guidelines demonstrates an awareness of how generative AI differs from earlier forms of AI in terms of capacity but carries over earlier assumptions about AI's limitations in legal judgment. In the fall of 2023, the Treasury Board released its *Guide on the use of generative artificial intelligence* for federal institutions.[51] Noting the challenge posed by generative AI's "limited transparency and explainability," which may conceal bias or errors, the *Guide* proposes various best practices for mitigating this.[52] These include "formulat[ing] prompts to generate content that provides holistic perspectives," testing for bias, and opting not to use a tool if the risk cannot be managed.[53] The *Guide* is also alive to a key distinction: having a model to work on documents uploaded to it versus prompting it to draw on its training data alone (or that along with information gathered from a web search). More reliable output, the *Guide* asserts, can be obtained by "[using] grounding and prompt engineering so the models build responses from only the information you provide and control."[54]

Yet the *Guide* singles out administrative decision-making as the only use-case clearly to be avoided, and in doing so, it carries over earlier assumptions about AI's functional limits.[55] The stated concerns here are twofold. First, the "design and function of generative models can limit

49. *Ibid*, Appendix C, citing s 6.3.13.
50. *Ibid*, Appendix C, citing 6.2.3.
51. Treasury Board of Canada Secretariat, *Guide on the use of generative artificial intelligence* (Ottawa: TBS, first published September 2023, last modified 3 June 2025), online: <canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html> [perma.cc/5QVC-CLYJ] [Treasury Board, *Guide*].
52. *Ibid* under the heading "Challenges and opportunities."
53. *Ibid*, under the heading "Policy considerations and best practices."
54. *Ibid*.
55. *Ibid*, noting "generative AI may not be suited for use in administrative decision-making."

institutions' ability to ensure transparency, accountability and fairness" in decisions it makes or helps to make.[56] And using these tools to make "high-impact decisions" may violate a provider's terms of use.[57]

Concerns about violating contractual terms are important, but in ways to be explored in Part II below, they do not pose *technological* impediments to using generative AI in judgment. The more crucial point is that the *Guide* asserts rather than explains why generative AI is incapable of being used in ways that would mitigate fairness and transparency concerns in this context when it might be used that way in others. The *Guide* implies that generative AI should not be relied on to make high-impact decisions because it cannot do two things essential to making fair and acceptable judgments. It cannot decide a case in a manner perceived to be fair, due in part to the mystery of how it makes a decision (under the hood). And it cannot reason in the ways thought to be essential for legal judgment. Yet, in ways to be explored in Part II below, both assumptions are open to question.

Tribunals across Canada have drawn a similar red line against using language models to help draft or make decisions. A practice directive issued by Tribunals Ontario in early 2025 states that: "Adjudication is a human responsibility. Tribunal members hear cases and make decisions based on the evidence and submissions provided by parties. They do not use AI to write decisions or analyze evidence."[58] The claims here about humans being central to legal judgment and AI having no role to play in writing decisions rest at least in part on assumptions about AI's technical capabilities. The directive hints at this in its concession that "[t]he field

56. *Ibid*.

57. The Guide cites as examples OpenAI's prohibition on using their models for decisions about health, credit, employment conditions, law enforcement, and migration and asylum. It also cites Google's prohibition on using their tools in "domains that affect material or individual rights or well-being." See OpenAI, "OpenAI Usage Policies" (last modified 29 October 2025), online: <openai.com/policies/usage-policies> [perma.cc/G3BG-L89R]; Google, "Google Generative AI Prohibited Use Policy" (last modified 17 December 2024), online: <policies.google.com/terms/generative-ai/use-policy> [perma.cc/PW38-AHX8].

58. Tribunals Ontario, "Practice Direction on the Use of Artificial Intelligence (AI) in Tribunal Proceedings" (April 2025), online: <tribunalsontario.ca/documents/TO/Practice-Direction-on-AI_EN.html> [perma.cc/U85H-BQR6] [Tribunals Ontario, "Practice Direction on AI"]. The same language can be found in the Canadian Human Rights Tribunal, "Practice Direction: Use of artificial intelligence (AI) in Tribunal proceedings," online: <chrt-tcdp.gc.ca/en/about-us/practice-direction-use-artificial-intelligence-ai-tribunal-proceedings> [perma.cc/P5G6-6T93]. See also the Condominium Authority Tribunal of Ontario, "CAT Practice Direction: Use of Artificial Intelligence in CAT Cases" (Toronto: 1 December 2024), online (pdf): <condoauthorityontario.ca/wp-content/uploads/2024/11/Practice-Direction-AI-use-in-Tribunal-Proceedings.pdf> [perma.cc/B34D-XQ4V], noting at 3: "Tribunal Members are fully accountable for their decision-making. They do not use AI to analyze evidence, decide cases, or author decisions."

of AI is evolving rapidly" and that Tribunals Ontario will "continue to monitor its use and impact" and will "adjust [the directive] as necessary."[59] How AI's evolution might affect its role in judgment forms the focus of Part II of this paper.

II. *Generative AI's capabilities in judgment*

A question arising from the survey in Part I is whether large language models give rise to new capabilities that call into question earlier assumptions about AI's suitability for decision-making in high-impact cases in administrative law. If so, what are these new capabilities? And to what degree, if any, do they mitigate concerns about AI in relation to administrative law and justice? I proceed in this Part by first canvassing experiments using large language models to produce outlines of a decision in an appellate court case and then consider what the results imply in the context of administrative law and justice.

I should note at the outset, however, that a more pertinent experiment would be one that tests AI's ability to make the kind of decision a tribunal might make—drawing on materials typically submitted by the parties. I have opted not to attempt this for two reasons. What might be called first-order administrative decisions tend to vary so much in terms of scope, procedure, fact, and law as to rule out the possibility of finding a representative tribunal or set of them. Some tribunals, for example, adjudicate between civilian parties such as a landlord and tenant or an employer and union, whereas others adjudicate between an administrative official such as an investigating officer and a civilian (for example, the Environmental Protection Tribunal of Canada or a superintendent of motor vehicles), and still others decide matters only in response to a civilian applicant, such as a visa officer in immigration. There is a wide variety here in the amount of fact finding done and the way in which it is done: based on written submissions, oral hearings, or investigations.

The possible use to be made of a language model in each of these cases is thus situation specific—in ways that would detract from a larger set of observations about capacity that I seek to demonstrate here. The aim in what follows is instead to demonstrate a language model's core capabilities in legal reasoning and judgment when drawing on party materials in an adversarial forum. It is AI's evolving capability that forms the focus of the experiment, with implications for administration tribunals canvassed in the discussion that follows.

---

59. Tribunals Ontario, "Practice Direction on AI," *supra* note 58.

1. *Experiments using large language models in appellate judgment*

As noted in the introduction to this paper, a number of scholars have begun experimenting with using language models to draft legal decisions by uploading party materials to the model—in various contexts including appeal cases and arbitration.[60] In the interests of space, I briefly canvass an experiment conducted by an appellate lawyer, Adam Unikowsky, in the summer of 2024, that demonstrates a language model's ability to consistently render the outline of a decision in an apex court case that approaches the quality and sophistication of the actual decision of the court itself, along with the specific outcome.[61] Unikowsky did this by uploading to Anthropic's Claude 3.0 language model briefs from 37 of the United States Supreme Court cases in the current term and asking it to briefly outline a decision in each one. In earlier research, I have sought to replicate part of the experiment in relation to recent decisions of the Supreme Court of Canada.[62] I draw here on findings of both experiments to show that language models can be used in a way that overcomes many of the concerns about limitations of AI in legal judgement noted above, including normative judgement, contextual sensitivity, and selecting and applying relevant factors. In what follows, I provide only a brief summary of the nature and findings of these experiments,[63] before considering what they might mean for administrative law.

Unikowsky uploaded to Claude between three and five of the actual briefs filed in each case and prompted Claude to provide an outline of a decision of the US Supreme Court in three to four paragraphs.[64] He found that Claude decided 27 of the 37 cases the same way the Court did and that in the remaining ten, he "frequently was more persuaded by Claude's analysis than the Supreme Court's."[65] His write-up of the experiment includes a detailed discussion of six decisions rendered in the previous week.[66] 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,

60. Unikowsky, "Trust," *supra* note 10; Unikowsky, "Trust Part II," *supra* note 10; Gandall, Kieffaber, & McLaren, *supra* note 10; Posner & Saran, *supra* note 10; Diab, "Evolving Role of AI," *supra* note 10.

61. The experiment, conducted by Adam Unikowsky, is outlined in two online posts: Unikowsky, "Trust," *supra* note 10, and Unikowsky, "Trust Part II," *supra* note 10.

62. Diab, "Evolving Role of AI," *supra* note 10.

63. For a more detailed discussion, see Unikowsky, "Trust," *supra* note 10; Unikowski, "Trust Part II," *supra* note 10; Diab, "Evolving Role of AI," *supra* note 10. For a sceptical assessment of Unikowsky's experiment, and the inferences he drew from it, see James Grimmelmann, Benjamin LW Sobel & David Stein, "Generative Misinterpretation" (2026) 63:1 Harv J on Legis 229.

64. Unikowsky, "Trust," *supra* note 10.

65. Unikowsky, "Trust Part II," *supra* note 10.

66. *Ibid.*

which is hardly 'wrong.'"[67] 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.[68] To replicate part of the experiment, I uploaded to OpenAI's GPT 4.5 factums in *R v Singer*,[69] a case the Supreme Court of Canada heard in February of 2025 but has not yet decided (late 2025). Both Unikowsky's experiments and my own reveal three notable capabilities of language models in legal judgment.

First, by uploading into the model's context window party materials and prompting the model to rely on only those materials, the model is effective in carrying out many of the basic reasoning tasks in judgment that scholars were doubtful about in Part I of this paper. Each outline of a decision in both experiments contained, in the first paragraph or two, a brief but sufficiently nuanced and accurate summary of the relevant facts and the parties' positions.[70] 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 which, in the case of Unikowsky's experiment, closely approximated in quality reasons of the US Supreme Court itself. In the interests of space, I provide the full result of only my experiment with *Singer*.

R v Singer is a case about whether police conducted an unlawful search when they came onto a private driveway to investigate a report about an impaired driver and found the driver asleep in his vehicle.[71] I uploaded to GPT-4.5 nine of the factums filed at the Supreme Court of Canada in the case and prompted it: "Drawing only on these factums, write the outline of a decision of the court hearing this appeal. The outline should be 4 to 5 paragraphs. Decide whether police violated section 8 of the Charter in this case and whether the evidence should be excluded."[72] The model produced an "Outline of the Court's Decision" in favour of the defence. I then asked the model to decide the case the other way, and

67. *Ibid*, referring to *Campos-Chaves v Garland*, 602 US 447 (2024).
68. 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" (2025) 100:1 Chicago-Kent L Rev 141.
69. *R v Singer*, SCC case 41090 [*Singer*].
70. For extended excerpts from Unikowsky's experiment, see Unikowsky, "Trust," *supra* note 10; Unikowsky, "Trust part II," *supra* note 10. I set out the result of my experiment with Singer, *supra* note 69, in full below.
71. *Singer, supra* note 69.
72. I refer in this prompt to section 8 of the *Canadian Charter of Rights and Freedoms*, Part I of the *Constitution Act, 1982*, being Schedule B to the *Canada Act 1982* (UK), *1982*, c 11.

it produced a similar outline with cogent and persuasive reasons for the opposite holding:

> 1. **Overview:** The central issue is whether police conduct violated Wayne Lester Singer's right under section 8 of the Canadian Charter of Rights and Freedoms, protecting against unreasonable searches and seizures. The appeal concerns police entering onto private property without a warrant to investigate a vehicle following a complaint of impaired driving. The Court must determine whether the officers' actions amounted to a lawful search and, if any Charter violation occurred, whether the evidence obtained should be admitted under section 24(2).
>
> 2. **Analysis of Section 8 Violation:** The vehicle was parked and running in a private driveway, visible from the public road, following an impaired driving complaint. Police officers approached the vehicle with the intent to investigate the ongoing complaint and observed Mr. Singer unresponsive in the driver's seat. The implied licence doctrine, permitting limited entry onto private property for investigative purposes and communication, applies here. Given the officers' legitimate investigative purpose, the intrusion was minimal and reasonable under the circumstances. Thus, the police conduct did not constitute an unreasonable "search" within section 8.
>
> 3. **Reasonableness of the Search:** The Court finds a reduced expectation of privacy regarding vehicles, even when parked on private driveways, especially under circumstances involving potential danger to public safety such as impaired driving. Given the urgency of ensuring public safety and the vehicle being clearly visible and running, the officers' actions in approaching and inspecting the vehicle were reasonable. The police acted appropriately, balancing individual privacy rights with society's interest in protecting against impaired driving.
>
> 4. **Admission of Evidence Under Section 24(2)**: Even if a minor infringement occurred, it was neither deliberate nor egregious. The officers acted in good faith, responding appropriately to an urgent public safety concern. The impact on Mr. Singer's Charter-protected interests was minimal, given the nature and location of the intrusion. Additionally, society has a compelling interest in the effective adjudication of impaired driving offences, a serious threat to public safety. Consequently, excluding evidence that was collected under circumstances where officers acted reasonably would bring the administration of justice into disrepute. Therefore, the evidence should be admitted.

GPT 4.5's output in my *Singer* experiment cannot at present (late 2025) be compared with the Court's final decision. However, the draft decision the model produced for *Singer* was clearly responsive to the arguments set out and issues raised in the factums, in a similar way in which decisions

in Unikowsky's experiments were responsive to the factums he uploaded to Claude.[73]

The larger point here relates to a language model's ability to outline facts and issues, and to decide a case with clear and cogent reasons. This demonstrates that generative AI—when prompted to rely on party materials—can be sensitive to social context (lower privacy expectations in a driveway where safety is at issue), can engage in normative reasoning (police should *not* require a warrant here) and can decide a case by applying relevant factors.[74] Or rather, a language model can *appear* to do so. I set aside for the moment the significance of this reality-versus-appearance distinction and return to it below. After the model produced the first "Outline of the Court's Decision," as noted, I followed up by asking the model to a generate the contrary holding cited above.

A second notable finding is that large language models can formulate new legal tests, ones not found in the materials themselves. This is not a task many administrative tribunals are asked to carry out. But it is relevant here to illustrate that language models can engage in what appears to be creative legal thinking. Unikowsky explores this capability in some depth,[75] but in the interests of space, I provide an example from my own experiment with *Singer*. After the model produced the "Outline of the Court's Decision" cited above, as noted, I followed up by asking the model to a generate a contrary holding. What GPT 4.5 did not do—without having been explicitly prompted to do so—is incorporate into its draft decision for either outcome a response to an invitation to formulate a new police power. Three of the factums had done so (those filed by the Attorneys General of Saskatchewan, Ontario, and British Columbia). That is to say, the model did not respond to the argument that the Court should craft *a new legal test* for when police are authorized to come onto residential

73. Unikowsky, "Trust," *supra* note 10; Unikowsky, "Trust Part II," *supra* note 10.

74. We might query whether privacy issues have a normative structure that makes it easier for language models to simulate normative judgments in distinction to other more multidimensional and contested constitutional issues. On one reading, a language model might excel at dealing with privacy questions but do less well dealing with conflicts between rights such as liberty and equality. This is a question worth pursuing through further experimentation. But it is notable here that Unikowsky's experiments with 37 cases before the US Supreme Court involved a host of constitutional questions with normative dimensions. His finding that Claude's output closely tracked the Court's actual output in many of these cases suggests that normative questions aside from privacy may not pose much of an impediment for a language model. I credit one of the anonymous reviewers of this paper for raising this point.

75. See his discussion, in Unikowsky, "Trust Part II," *supra* note 10, of testing Claude 3.0's ability to come up with successively more creative and elaborate versions of a legal test in *Lindke v Freed*, 601 US 187 (2024) for deciding whether First Amendment rights are engaged in response to statements by public officials in non-governmental settings.

property in a case like this to conduct a search. I note that only Ontario's factum sketched the specific wording of a test. I prompted GPT 4.5:

> In this second outline, deciding in favour of the prosecution, you found the implied license doctrine permitted entry for limited purposes. Produce another outline of a decision in the prosecution's favour, but this time make the section 8 analysis turn on the Court's recognition of a new ancillary-police power to carry out a limited search of a vehicle on a private driveway in impaired driving cases and briefly set out the test for this.

The main paragraph of the model's new outline contained an entirely new test:

> The Court recognizes a new ancillary police power permitting a limited search of a vehicle on private property in specific circumstances involving impaired driving. The test for this ancillary power is as follows: (1) the officers must have reasonable grounds to suspect that impaired driving has occurred; (2) the vehicle must be readily accessible or visible from public space; (3) the intrusion must be minimal and proportionate to the suspected offence; and (4) there must be a genuine concern for public safety requiring immediate investigative action.[76]

While there was a basis for portions of this in the factums—reasonable suspicion as grounds, exigency—the test that GPT 4.5 proposed does not map closely on to Ontario's proposed test or to arguments set out in the other two factums.[77] The Saskatchewan Court of Appeal's decision in *Singer* also only treats the ancillary powers doctrine briefly, dismissing the argument that it should apply here.[78]

A third notable finding—directly applicable in the administrative law context—is that language models can apply legal tests to novel facts, demonstrating an ability to decide a case on the basis of relevant factors and those alone. An example can once again be found in my experiment with *Singer* and GPT 4.5's response to the prompt set out above asking it to draft a decision containing a new legal test. Without being explicitly prompted to do so, the model's draft decision went a step further (after the paragraph cited above) and applied the new test to the facts in this case:

> Applying this newly recognized ancillary power to the present case, the officers had reasonable grounds based on the complaint received

76. ChatGPT, "Section8 Decision Outline" (OpenAI, 2025) <chatgpt.com/canvas/shared/682c92a4ff08191a9d8ee98401f2222> [output on file with author].

77. See the factums filed in *Singer*, *supra* note 69, online: <scc-csc.ca/cases-dossiers/search-recherche/41090/> [perma.cc/VDL8-6EYC] [*Singer* Factums].

78. *R v Singer*, 2023 SKCA 123, dismissing the ancillary powers submission at para 68.

> and the observable condition of the vehicle running on the driveway. The vehicle was visible and easily accessible from the street, and Mr. Singer was unresponsive, indicating a potential risk to public safety. The intrusion—opening the vehicle door to assess Mr. Singer's condition—was minimally invasive and proportionate given the circumstances. Consequently, the search was reasonable under section 8 of the Charter.[79]

In this brief passage, GPT 4.5 demonstrates an ability to single out key facts and to apply relevant legal factors in an assessment of them: reasonable grounds, risk, proportionality. This demonstrates a model's ability to discern which facts are relevant to certain legal issues, and its ability to confine its decision to those factors alone.[80]

2. *Implications for administrative law*

What broader inferences can be drawn from this experiment for using AI to make decisions in administrative law? There are similarities in some of the forms of legal reasoning found in both appellate courts and in administrative tribunals, such as identifying material facts and issues, issues, and applying legal tests. But the materials submitted to a tribunal are often brief or strictly factual. Whether a language model could produce a draft of reasons for decision in an administrative law context would thus depend at least in part on the nature of the material typically provided to a given tribunal or decision-maker. For this reason, as noted above, a more fulsome assessment of a language model's capabilities in administrative judgement would require a similar set of experiments to be conducted using material typically submitted to a given tribunal, such as a human rights commission, or an employment standards or labour board.[81]

However, the findings with respect to *legal reasoning and judgment* capabilities in the experiments canvassed here are likely to be the same in the tribunal context where a language model is provided with a sufficient body of core material. Any of the frontier language models at present (GPT 4.5, Claude 4, etc.) could be prompted to draw exclusively on documents filed by the parties, along with a relevant statute, a tribunal's rules of process, and possibly a set of key principles from the tribunal's earlier decisions.

---

79. The full text of the output to this prompt can be found at the link in "Section8 Decision Outline," *supra* note 76, in paragraph 3 of the output.

80. A further example of this can be found in GPT 4.5's application of s 24(2) in its original output siding with the prosecution in *Singer*, *supra* note 69.

81. As noted, Paul Daly has embarked upon this in relation to the IRCC in Daly, "Artificial Intelligence," *supra* note 9 at 15-18, prompting GPT with facts and law but not providing it with documents submitted by the parties.

To be clear, the argument here is not that a language model can replace a decision-maker in every administrative context or automate every aspect of administrative adjudication. There are some things that only humans can do—or do well enough—at present and for some time to come: make findings of fact, assess credibility, and manage the conduct of the parties before the tribunal. And it may be the case that only a small few tribunals receive materials analogous to the detailed, well-argued factums used in the experiments above—calling into question the relevance of those experiments for tribunal judgment. The response is to emphasize that the argument here is more limited: administrative *decision-making*, where it draws in large part on a discrete body of material (rather than findings of fact, credibility), does not entail a set of abilities beyond the scope of large language models at present (or many earlier assumptions about AI's limits in this regard are no longer safe to assume). This means that in at least some instances, where tribunals do receive more fulsome materials, and decisions in the context at issue turn mainly on applying law to facts not contested, language models could, in theory, be capable of making decisions by drawing on them.

Even if this is conceded, however, there might still be concerns about control, privacy, and security. In terms of control, two issues with using language models to make decisions arise. One relates to the use of "Retrieval Augmented Generation," (RAG), or the technique that models employ to draw on an external body of documents, i.e. party materials, such as factums and briefs.[82] RAG has helped models render more accurate output by loading content or information from a document into the model's context window to make it part of the prompt. But this does not completely overcome a model's tendency to hallucinate, with evidence pointing to the persistence of errors in AI that employs RAG in legal databases.[83] How much a model hallucinates when drawing on a small number of factums is less clear, but it is a concern worth noting and mitigates against full reliance on AI for judgment without human oversight. A second set of concerns arises from the fact that language model judgments are known to be "highly sensitive to prompt phrasing, output processing methods, and model training choices," leading to different outcomes and varying degrees of confidence in them based on slight changes in the wording of a

82. Ivan Belcic, "What is RAG (Retrieval Augmented Generation)?," online: <ibm.com/think/topics/retrieval-augmented-generation> [perma.cc/UJ97-3QY3].
83. Magesh et al, "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools" (2025) 22:2 J Empirical Leg Stud 216.

prompt.[84] Researchers have proposed technical solutions to these issues,[85] but this concern also mitigates against full reliance on AI without human oversight.

In terms of privacy and security, the experiment here demonstrates what large ("frontier") language models can do, but smaller, bespoke models created for a specific tribunal might do just as well. This may be relevant where a tribunal seeks to keep aspects of the adjudication confidential. To ensure data privacy and security (and to comply with terms of use), a tribunal or agency may not be in a position to rely on a large provider such as OpenAI or Google. Yet the technology here is quickly evolving to address this issue, with smaller models that one can operate offline becoming more powerful.[86] A smaller open-source model could conceivably be modified for a tribunal's purpose and located in-house.[87] An analogous experiment would then need to be run in a given context to test the proposition that a smaller, off-line model could produce an administrative decision of comparable quality to that of a human. There is, however, no clear reason at this point to assume that this is not possible. Privacy and security concerns about using generative AI in sensitive contexts are not insurmountable barriers.

To return to the main point, if a tribunal can provide sufficient material to a language model, the findings of the experiments canvassed here suggest that a model could generate a draft of a reasoned decision in many administrative contexts. Large language models are capable of producing reasons of a sufficient quality to meet the test for reasonableness in *Vavilov*.[88] That is to say, models can generate reasons that provide a cogent and transparent justification for a decision, one that is responsive to both law and fact and does not consider incorrect or improper factors. Further experiments would need to be done to confirm that this also applies to smaller, in-house models.

---

84. Jonathan H Choi, "Off-the-Shelf Large Language Models Are Unreliable Judges," (2027) J Empirical Leg Stud (forthcoming), online: <papers.ssrn.com/abstract=5188865> [perma.cc/FMD5-KP3D].

85. Jack Kieffaber et al, "LLMs are Bad Judges. So Use Our Classifier Instead" (2025) 78 SMU L Rev Forum 57.

86. In early 2025, for example, a Chinese AI company Deep Seek purported to have created a language model with comparable capacity to the various frontier models at a fraction of its size. See Zeyi Yang, "How Chinese AI Startup DeepSeek Made a Model that Rivals OpenAI," *Wired* (25 January 2025), online: <wired.com/story/deepseek-china-model-ai> [perma.cc/842U-5MKM].

87. John Johnson, "Small Language Models (SLM): A Comprehensive Overview" (22 February 2025), online (blog): <huggingface.co/blog/jjokah/small-language-model> [perma.cc/4UV4-XUXR].

88. *Vavilov*, s*upra* note 16. This assumes that questions of underlying opacity do not on their own form a basis for finding an AI decision unreasonable, a point canvassed in Diab, "*Vavilov* and Generative AI," *supra* note 18.

3. *Broader issues with using generative AI in administrative decisions*

I assume in this final segment that a language model can outline a decision in an administrative law context and provide reasons to support it that rival in quality those of a human. Should this form of AI be used to make decisions alone in high-impact cases? Where a human is tasked with deciding case, if they prompt a language model to draft a decision with reasons, would this entail an improper subdelegation? Where and how do we draw the line between deciding and assisting? Once again, earlier administrative law scholars have canvassed these questions[89]; in what follows, I explore them briefly in light of the new capabilities surveyed above.

a. *Should language models be used to make high-impact decisions alone?*

Scholars have offered three core arguments against AI making decisions alone in high-impact administrative law cases. New capabilities point to a response in each case.

The first argument is that AI—including language models—makes decisions based on machine learning processes that are opaque and may conceal "embedded systemic bias."[90] As noted earlier, language models are so complex as to render their working explainable in a general sense but not interpretable in a specific sense.[91] This fundamental limitation precludes any assurance that AI's output is not compromised by bias in the algorithms that shape the output, or that a model's inner workings cannot be gamed in some way to engineer an output.[92] These facts alone, or concerns about them, render generative AI just as unsuitable for full reliance in high-impact administrative decision-making as earlier, more limited forms of AI canvassed above.[93]

One response is to require that a model be auditable or testable for bias and to insist on a measure of explainability as to its inner workings[94]—

89. Daly, "Artificial Administration," *supra* note 5; Daly, "Artificial Intelligence," *supra* note 9; Tan, *supra* note 42.

90. Daly, "Artificial Intelligence," *supra* note 9 at 23; Goudge, *supra* note 4 at 24; Beatson, *supra* note 4 at 308.

91. Arrieta et al, *supra* note 14.

92. *Ibid* at 83.

93. This is at least implied in the Treasury Board's *Guide*, *supra* note 51, in its caution against the use of generative AI to make administrative decisions, as noted in Part I above. See also Goudge, *supra* note 4 at 35.

94. Scholarship on language model testing for bias and on model explainability recognizes challenges and limits on both fronts but assumes a reasonable measure of effectiveness is possible. 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" (2026) 15:9 Electronics 1; see also Leif Azzopardi & Yashar Moshfeghi, "PRISM: A Methodology for Auditing Biases in Large Language Models" (last modified 10 November 2024), online: <arxiv.org/abs/2410.18906> [perma.

and if these conditions are met, apply a presumption of impartiality to a model's output. A rule might be formulated which holds that where a model's inner workings are audited and reasonably explainable, its output should be presumed to be reliable—in the absence of evidence that AI's output reflects bias or has been gamed.[95] This would shift the focus in most challenges to a language model's decision—as is the case with humans—to whether the decision and its reasons meet the test for reasonableness in *Vavilov*.[96]

A second argument against language models making decisions alone here also relates to the technical basis of the decision. Amin Afrouzi argues that, under the hood, generative AI, like earlier forms of AI, decides a case in a fundamentally different way than a human does.[97] AI generates output by carrying out predictions based on correlations and patterns among words, not by consciously comprehending their meaning. These "syntactic" correlations are the true basis or rationale for any decision that AI makes.[98] AI does not decide a case on what he calls a "normative basis" or because a given outcome is what law or morality compels.[99] The reasons that a language model might provide for a decision in its textual output are merely ex post justifications for it. Humans might behave the same way, he concedes, deciding first upon an outcome and then coming up with reasons to justify it.[100] But, in theory, they are capable of deciding a case on a normative basis using "conscious human reasoning."[101] They can decide a case ex ante on the basis that it is the most or only morally correct outcome. And this rationale—and no other—is what we expect of a lawful judgment.[102]

---

cc/ZN85-NFQX]. On model explainability, see the survey in Haiyan Zhao et al, "Explainability for Large Language Models: A Survey" (2024), 15:2 ACM Transactions on Intelligent Systems and Tech 1.

95. Eugene Volokh, "Chief Justice Robots" (2019) 68:6 Duke LJ 1135 at 1168, contemplates a similar requirement in calling for AI judges to be testable in advance for bias, and if not audible then developed on the principle of "impartiality by design" (being "programmed to ignore certain attributes, such as parties' race, in drawing its generalizations" or having "training data…vetted to minimize bias flowing from that data.") See also Coglianese & Lehr, *supra* note 7 at 1216, envisioning a similar practice: "[administrative] agencies should increasingly seek out and engage neutral statistical experts to provide dispassionate assessments of consequential uses of algorithms."

96. *Vavilov*, *supra* note 16, suggesting at para 100 that a decision is presumed to be reasonable by virtue of the fact that "[t]he burden is on the party challenging the decision to show that it is unreasonable."

97. Amin Ebrahimi Afrouzi, "John Robots, Thurgood Martian, and the Syntax Monster: A New Argument Against AI Judges" (2024) 37:2 Can JL & Jur 369.

98. *Ibid* at 375-376. See also Coglianese & Lehr, *supra* note 41, for a similar argument.

99. Afrouzi, *supra* note 97 at 375.

100. *Ibid* at 379, 386-388.

101. *Ibid* at 375, 387.

102. *Ibid* at 383: "…*legally valid* decisions must be correct not only in their holding but also in their

A possible response to this is that a language model's process for arriving at a decision (under the hood) may be purely statistical or correlative in nature—in ways that remain largely opaque to us—but the decision should still be acceptable if the textual reasons provided to justify it are cogent and persuasive. Anticipating further advances in legal AI, Eugene Volokh made this argument in 2019.[103] Where either a human or AI renders a decision, Volokh contended, the primary basis (though not the only one) on which to assess its legitimacy or acceptability is how persuasive we find its reasons. Human judges might notionally strive to arrive at a single right answer, one that is "fair, wise, or correct."[104] But in contentious cases, perspectives on this will differ. The only way that humans have to decide which decision is most correct—which is to say, the single most morally legitimate outcome—is to articulate reasons for it that seem most persuasive.[105] If persuasion is a primary basis of legitimacy in judgment, and an AI decision can be highly persuasive (as canvased above), the decision should be recognized as legitimate on this basis. (More on other sources of legitimacy below.)

Volokh's response extends to concerns raised in the first argument about bias or opacity in AI's inner processes. In the absence of evidence that an AI's process has been gamed or is biased, if the reasons AI provides stand up to human scrutiny, if they provide a clear, transparent, and persuasive justification for a given outcome that is legally correct (considers the right factors, applies the legal test correctly), then it should not matter what happens under the hood. This would be consistent with the holding in *Vavilov* on reasonableness. If reasons "justify to the affected party, in a manner that is transparent and intelligible,"[106] how a decision was arrived at—not technically, under the hood, but on the basis of fact and law—the underlying process should not matter here, just as it does not in the case of humans.

A third argument is that even if a language model can decide a case without violating principles of administrative law, having a machine making a high-impact decision could not satisfy a broader conception of

rationale." See also Williams, *supra* note 26 at 487, drawing a similar distinction: "where ML 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 recognizing that, statistically, that factor often correlates with the relevant outcome."

103. Volokh, s*upra* note 95

104. *Ibid* at 1152.

105. *Ibid* at 1153: "In 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."

106. *Vavilov*, *supra* note 16 at para 96.

administrative justice.[107] Administrative law scholars distinguish between principles that render a decision lawful (including procedural fairness, reasonableness) and those that render it just.[108] Administrative justice refers more broadly to "those qualities of decision making process that provide arguments for the acceptability of its decisions."[109] As Paul Daly notes, "[a]cceptability, in administrative justice terms, is about matching a decision-making process to the appropriate decision-making model."[110] Daly refers here to Jerry Mashaw's commonly cited mapping of basic models of administrative justice.[111] These range from decisions, at the lower end, made within a "bureaucratic rationality" model, which applies to forms of "program implementation" that mostly involve "information processing" (e.g. renewing a license), to decisions at the higher end made within a "moral judgment" model, where "conflict resolution" is carried out through the use of "contextual interpretation" and with expectations about fairness and independence.[112] While the bureaucratic model requires only accuracy, legitimacy within the moral judgment model rests on the "addition of a human touch to the decision-making process."[113]

In response to this line of argument, Volokh has suggested that what we perceive here as legitimate may be culturally or historically contingent.[114] He concedes the possibility that "people will never get used to AI judges" but contends there is "no reason to write off AI judging just because many people's first reaction to the concept may be shock or disbelief."[115] As generative AI becomes a more pervasive technology, the prospect of its use in high-impact judgment may not be associated so readily with "bureaucratic rationality." It might come to be perceived as an acceptable substitute for the human touch in the moral judgment model. An analogous shift currently unfolding can be found in the rising popularity of therapy bots.[116] A substantial public willingness to accept that a language model

107. Daly, "Artificial Administration," *supra* note 5 at 11-14.
108. Adler, "Understanding Administrative Justice," supra note 5, cited in Daly, "Artificial Administration," *supra* note 5 at X.
109. *Ibid*.
110. Daly, "Artificial Administration," *supra* note 5 at X
111. Jerry Mashaw, *Bureaucratic Justice: Managing Social Security Disability Claims,* (New Haven: Yale University Press, 1983), cited in Daly, "Artificial Administration," *supra* note 5 at X.
112. *Ibid*.
113. Daly, "Artificial Administration," *supra* note 5 at X.
114. Volokh, *supra* note 95 at 1171, noting, "[p]eople's eventual reaction to a new invention, after they are used to it, may be much friendlier than their initial reaction." He points to life insurance and in vitro fertilization as examples.
115. *Ibid*.
116. Anisha Sircar, "Could A Bot Be Your New Therapist? How AI Has Transformed Mental Healthcare," *Forbes* (28 October 2024), online: <forbes.com/sites/anishasircar/2024/10/28/could-a-bot-be-your-new-therapist-how-ai-has-transformed-mental-healthcare> [perma.cc/8NMH-4GVF].

can substitute for a therapist (if not a romantic companion)—*despite the many concerns this raises*—supports the inference that automated processes can gain acceptance and be associated with something other than "bureaucratic rationality."[117]

One possible solution to concerns raised with all three arguments is to rely on language models to make decisions alone in high-impact cases on the condition that both parties consent, with the option of review by a human judge at the behest of either party.[118]

b. *To what degree can humans rely on language models to make decisions without delegating?*

What about situations where humans are tasked with deciding a case? As Paul Daly has noted, there would be no issue with a person prompting a language model to outline reasons for a decision that they had already arrived at.[119] He compares this possibility to asking a clerk to draft an opinion for a given outcome.[120] The question is whether placing any greater reliance on AI would violate the principle that "he who hears a case must decide."[121] Does a decision-maker delegate their discretion if they first ask a language model to decide and then merely affirm or ratify the model's draft opinion? Arguments about automation bias would raise the concern that in this case, it is the machine rather than the person who makes the decision.[122] Would this concern be overcome—would a human delegate their discretion—if they were to ask a language model to produce outcomes for and against either party and choose among them?

Daly points out that the problem of the "lazy decision-maker" is not new with AI; with techniques such as cutting and pasting, there have long been ways of cutting corners.[123] His recommendation in the case of AI for avoiding delegation to a language model would be to adhere to the

117. Anna Mae Duane, "Teenagers Turning to AI Companions Are Redefining Love as Easy, Unconditional and Always There," *The Conversation* (12 February 2025), online: <theconversation.com/teenagers-turning-to-ai-companions-are-redefining-love-as-easy-unconditional-and-always-there-242185> [perma.cc/Q5JJ-GULZ].

118. Unikowsky in "Trust," *supra* note 10, proposes these solutions for using AI in appellate judgement, suggesting that it would seem desirable to many parties for offering a quicker means of reaching a resolution.

119. Daly, "Artificial Intelligence," *supra* note 9 at X.

120. *Ibid*.

121. *IWA v Consolidated-Bathurst Packaging Ltd*, [1990] 1 SCR 282, 1990 CanLII 132 [*Consolidated-Bathurst*].

122. See Smith and Brown, *supra* note 25. The Supreme Court of Canada in *Consolidated-Bathurst*, *supra* note 121 at 331, has, however, rejected the proposition that "any discussion with a person who has not heard the evidence necessarily vitiates the resulting decision because this discussion might 'influence' the decision maker."

123. Daly, "Artificial Intelligence," *supra* note 9 at 21.

general principle that "the greater the impact of the decision in question, the more closely the decision-maker's engagement with the material ought to be scrutinized."[124] Where avoiding subdelegation to AI remains a priority, Daly's insight helps to illuminate where the line is to be drawn (i.e. substantial versus cursory engagement with the material).

Yet as AI becomes more pervasive, perceptions about the propriety of subdelegation may shift, and where the line is to be drawn between reliance on a language model and delegation of a decision to it may become less clear. It seems conceivable at this point that legislation could permit language models to be used in high-impact administrative decision-making in ways that blur the line between assistance and delegation. This might be done, for example, by allowing a tribunal to use AI to draft reasons or, under certain conditions, to merely affirm decisions that AI has drafted. Courts would need to decide the acceptability of these possible forms of subdelegation as consistent with procedural fairness in those contexts, but perceptions here too might also change in step with shifts in the broader culture.

Conclusion

This paper has argued that language models offer new ways in which to use AI that effectively automate core processes in administrative decision-making and in legal judgment more broadly. Earlier scholarship and policy had assumed that AI would likely be confined to an assistive role in administrative law based on functional limitations reflected in earlier forms of AI. Yet even after the advent of generative AI, assumptions about AI's functional limits have continued to shape law and policy.

Although scholars and jurists have begun to consider the possible uses of language models to assist in formulating decisions, experimentation with the capabilities and affordances of generative AI is only beginning. The experiments in this paper focused on a language model's ability to decide a case and provide cogent and persuasive reasons when prompted to work on party materials in appellate cases. The paper extrapolated from this a language model's ability to decide cases in administrative law contexts where findings of fact and relevant legal tests are included among the material a model is prompted to consider.

The paper then canvased objections to AI playing a larger role in administrative decision-making. These pertained to principles of both administrative law and justice. Two key shifts would need to take place in sociocultural perspectives on the use of AI in judgement. Legislatures,

124. *Ibid*.

courts, and tribunals would need to embrace the proposition that the appropriate bases on which to assess the validity of an automated decision are the cogency and persuasiveness of its reasons for judgment—not whether AI is capable of making a decision on a normative rather than purely syntactic or correlative basis, or whether a language model's process is fully interpretable and proven to be completely free of bias. More broadly, lawmakers and the wider public would need to accept automated decision-making in high-impact cases as a legitimate model of administrative justice.

Yet as language models become more capable and as administrators become more adept at using them, current law and policy envisioning the containment of AI to an assistive role in judgment are likely to be questioned. Objections to using AI on grounds of incapacity will likely be challenged, and with time, concerns on moral grounds may become less salient.