The Evolving Role of AI in Legal Judgment

Robert Diab, Professor, Faculty of Law, Thompson Rivers University

June 2025

Forthcoming in *Law, Innovation, and Technology* in 2026

Abstract:

In the 2010's, scholars had reached a consensus that the role of artificial intelligence in legal judgment was constrained by limitations inherent in machine learning and other forms of AI. AI could not reason by analogy, draw on a wider context, be sensitive to quickly shifting social norms, or effectively address algorithmic bias and opacity. But the AI critiqued here mainly involved tools for predicting scores or probabilities of an outcome. This paper argues that AI involving language models fundamentally challenges these earlier assumptions, calling for a reassessment of the prior consensus. Unlike earlier forms of AI, language models are increasingly effective at engaging in analogical reasoning, being context aware, and mitigating bias and opacity in judgement. The paper demonstrates this by canvassing an experiment using briefs and factums from apex court cases to render decisions with these models that closely match the actual outcomes in those cases. The findings point to legal judgment becoming increasingly collaborative between humans and machines, and distinctions between automated processes and human intuition becoming increasingly blurred.

Introduction 2

I. The earlier consensus on AI's role in judgment 3

II. Generative AI in judgment and mediation 16

a. Experiments in appellate judgment 17

b. Observations and implications 25

Reservations and conclusions 33

Introduction

Before the appearance of ChatGPT in 2022, scholarship on the role of artificial intelligence (AI) in judgment in courts of law and other forms of adjudication was extensive.[1] By the end of the 2010s, a consensus had emerged that AI could not replace human decision-making in adjudication due to a host of basic limitations inherent in the technology of machine learning and other forms of AI. While it could recognize patterns and make statistical predictions, AI could not reason by analogy or make normative judgments.[2] It could not assess wider contexts or draw on experience in making findings of fact;[3] it could not identify recent or quick changes in social norms;[4] and its processes were fraught with algorithmic bias and opacity.[5]

The scholarship prior to 2022, however, was largely premised on a narrow conception of AI as a tool for predicting scores or probabilities of an outcome. Paradigm cases for using AI in judgment included software that assessed the probability of a person reoffending or the likely finding of a court on a given question (employee or contractor?) based on similarity to a precedent.[6] The argument of this paper is that much of that literature — which continues to be cited as current — should be reassessed in light of the advent of generative AI involving transformer-based language models such as ChatGPT, Claude, and Gemini. This adaptation of machine learning to analyzing and outputting natural language makes possible a host of other capabilities in judgment beyond merely scoring or predicting probabilities. Since becoming widely accessible in 2022, generative AI has gained vastly greater capabilities. These facts alone call for a reassessment of what role AI can — not should but can — play in legal judgment.[7]

---

[1] An overview of the literature can be found in Part I below.

[2] Harry Surden, 'Artificial Intelligence and Law: An Overview' (2019) 35(4) *Ga St U L Rev* 1305; Cass R Sunstein, 'Of Artificial Intelligence and Legal Reasoning' (2001) 8 *U Chicago L Sch Roundtable* 29.

[3] John Morison and Adam Harkens, 'Re-engineering Justice? Robot Judges, Computerized Courts and (semi) Automated Legal Decision-Making' (2019) 39(4) *Legal Studies* 618.

[4] Tania Sourdin, 'Judge v Robot? Artificial Intelligence and Judicial Decision-Making' (2018) 41(4) *UNSWLJ* 1114.

[5] Ignacio N Cofone, 'Algorithmic Discrimination is an Information Problem' (2019) 70(6) *Hastings LJ* 1389; Jenna Burrell, 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) *Big Data Soc* 1.

[6] This can be gleaned from the examples canvassed in the literature in Part I below.

[7] The normative question of whether and to what extent automated processes should come to replace humans in the practice of judgment is a distinct issue and, although touched on throughout, is not the focus here. The aim in

This paper explores AI's wider capacity in judgment by canvassing, replicating, and expanding upon an experiment first conducted in 2024 that has begun to garner scholarly attention.[8] The experiment involves uploading briefs or factums in actual apex court cases to produce an outline of a decision. The results in the various test cases match not only the court's actual holding (in cases from the Supreme Courts of the United States and Canada) but also the court's main reasons for judgement, closely approximating them in quality and substance. The experiment shows that by drawing primarily on party materials, language models challenge earlier assumptions about AI's inability to reason by analogy, formulate or apply normative principles, or discern rapidly shifting social values. The results also show how AI used in this way can mitigate to significant degree concerns about algorithmic bias and opacity through the provision of detailed reasons for judgment that offer a clear basis for assessing the merits of the decision.

Part I of this paper provides an overview of the consensus among scholars prior to 2022 as to AI's limited capabilities in legal judgment — to articulate the nature of common concerns, but also how they related to the earlier state of the technology at the time. Part II canvases and expands upon the experiment with generative AI and apex court decisions noted above, followed by a discussion of what it implies about claims in the earlier literature and why concerns about opacity persist. The paper concludes by suggesting that while important gaps remain in AI's capabilities for adjudication at the trial level, its present abilities point to the likelihood of greater collaboration between humans and machines in judgment, with humans coming to play a more editorial, supervisory role. This in turn points in the direction of a further blurring of the distinction in judging between automation and human intuition.

I. The earlier consensus on AI's role in judgment

In an important contribution to an edited collection on AI in 2022, Ignacio Cofone offered a comprehensive overview of the scholarly consensus on the role of AI in judgment reached by

---

this paper is to canvas the antecedent question of whether AI *can* perform tasks considered essential to legal judgment, and how well it can or cannot do them.

[8] The experiment, conducted by Adam Unikowsky, is outlined in two online posts: 'In AI we trust', June 8, 2024, <https://adamunikowsky.substack.com/p/in-ai-we-trust> accessed 7 April 2025, and 'In AI we trust, part II', June 16, 2024, <https://adamunikowsky.substack.com/p/in-ai-we-trust> accessed 7 April 2025. A brief discussion of it can be found in André Guskow Cardoso and others, 'Generative Artificial Intelligence and Legal Decisionmaking' (2024) 19(11 and 12) *Global Trade and Customs Journal* 710, 711-12.

the end of the 2010s.[9] The irony one senses when reading Cofone's excellent overview in hindsight arises from knowledge that the landscape would soon rapidly shift with ChatGPT's public release later in 2022, and the race that has since unfolded to develop generative AI. While Cofone's chapter contains what is still an indispensable survey of a debate stretching over a decade about the limits of AI in adjudication up to that point, it also now stands as an artefact from an earlier era in technology. I draw on Cofone's survey of the terrain to lend a clearer picture of what was commonly assumed, prior to 2022, about the scope of AI's possible role in judgment, given its various limitations. In ways to be explored, many of the assumptions framing this picture will turn out to have rested on limits in the technology at that earlier stage in time.

We begin to glean this connection from the fact that in making arguments about AI's potential role in judgment, scholars have specific AI tools in mind. Cofone's own paradigm cases for AI's use in judgment include COMPAS, the software used in US courts to predict the probability of violent re-offending; TrueAllele, a program for analyzing incomplete DNA samples; and software the Canadian tax authorities use to assess the likelihood of someone being a tax evader.[10] In each case, the AI tool is limited to measuring the probability of an outcome — generating a score — based on a degree of similarity between a set of variables old and new.

While generative AI would soon unsettle this conception what AI can do, much of the literature on judgment that Cofone cites unfolds within a context in which limited purpose tools such as COMPAS shape assumptions about the scope AI's capability. This context serves as an unspoken qualification to the general thrust of Cofone's survey: that a consensus had emerged at that point in time to the effect that while AI could assist judges with predictions or pattern matching tasks, there remain many facets of legal judgment that only humans can carry out. With predictive tools becoming more capable, AI was 'increasingly likely to assist judges in their decision-making,'[11] but it would assist them in limited ways. Before considering the question of whether and how newer AI unsettles this consensus view, I canvas in more detail some of the authorities that Cofone cites to support this general position.

---

[9] Ignacio N Cofone, 'AI and Judicial Decision-Making' in Florian Martin-Bariteau and Teresa Scassa (eds), *Artificial Intelligence and the Law in Canada* (Toronto: LexisNexis Canada, 2021) SSRN: <https://ssrn.com/abstract=3733951> accessed 7 April 2025 (citing to SSRN version).

[10] Ibid 4.

[11] Ibid.

Cass Sunstein argued in 2001 that AI was, at least at that time, confined to pattern recognition whereas legal judgment involves reasoning by analogy, which in turn involves identifying and applying normative principles.[12] Put another way, a machine can assess a degree of similarity between cases based on the number of similar variables. But a machine cannot decide which variables are *relevant* for making a comparison. This decision involves a value judgment.[13] And this, in turn, depends on identifying the normative principle on which to base a comparison.[14] Two cases cannot be said to be similar in a relevant or material way unless one decides that each stands for the same underlying principle: *e.g.*, *TikTok v Garland* (challenging the ban on the app in the United States) is really about the need to protect a company's right to free expression from undue state infringement, or it's really about the need to defer to government limits on a company's speech so as to protect against real threats to national security.[15] Until one identifies this principle — one for which an earlier set of cases stands — no *degree* of similarity with those cases will matter. In short, AI only works in legal judgment when humans frame a given case under a broader normative principle, or when they decide what it is fundamentally about or how it should be framed. And in the early 2000s, Sunstein concluded, AI could not do this on its own.[16]

Another line of argument Cofone outlines involves skepticism about machines being sensitive enough to the wider social context and human dynamics in which judgment unfolds. John Morison and Adam Harkens argue that 'the practice of law, and the role of judges, is fundamentally socially produced and acted on by dynamic processes within the wider legal system which are complex, and contingent on a social context in ways that it is difficult to

---

[12] Ibid, citing Cass Sunstein, 'Of Artificial Intelligence and Legal Reasoning' (2001) 8 U. Chicago L. Sch. Roundtable 29, 29 and 33-34.

[13] Sunstein ibid 31.

[14] Ibid 31-32.

[15] The Supreme Court in *TikTok v Garland*, 604 US ___ (2025) agreed that the act of Congress compelling ByteDance's divestiture of the platform in the United States had engaged the company's First Amendment rights. But it held, at 12-13, that a lower standard of justification for infringing speech rights of the company was appropriate here and was met because 'TikTok's scale and susceptibility to foreign adversary control, together with the vast swaths of sensitive data the platform collects, justify differential treatment to address the Government's national security concerns. A law targeting any other speaker would by necessity entail a distinct inquiry and separate considerations.' It is, in short, *better understood* as an instance of earlier national security cases than earlier free speech cases.

[16] Sunstein's conclusions were nuanced, noting, at 34 (n 12), that '[c]omputer programs do not yet reason analogically. But this proposition should not be confused with the suggestion that in the nature of things, evaluative arguments are uniquely the province of human beings, or that computer programs will never be able to help human beings with it, or even to engage in it on their own.'

imagine [computers] capturing in full or accurately.'[17] Judgment and adjudication involve many elements that machines are 'some way off replicating the human umpire.'[18] These include issue identification, fact finding, 'ascertaining the relevant legal framework, providing an opportunity for venting feelings, evaluating the parties' interests, disaggregating issues, establishing positions, exchanging information, [and] suggesting options for resolution'.[19] Given the extent to which much of this requires an attentiveness to social context, the use of discretion, or emotional intelligence, it seems unlikely that these elements in legal processes can be taken over by an 'algorithmic actor'.[20]

Tania Sourdin and Richard Cornes have argued in a similar vein that 'the role of the judge is multifaceted.'[21] It can involve 'complex activism, complex interactions with people, dispute settlement, case management, public and specific education activities, social commentary as well the core adjudicatory functions'.[22] These in turn draw on a judge's intuition, their 'curiosity, emotional understanding of parties and their lawyers,' and their skill in 'agile questioning and exploration.'[23] More broadly, the authors contend, judges must be 'responsive to contextual factors that assist to determine the meaning of legislation and human activities.'[24] In novel situations, this raises a significant challenge for AI, since 'machine learning techniques are only useful where analysed information is similar to new information presented to the AI.'[25] Even if AI could surmount this hurdle, the 'social legitimacy' and 'respect for judicial judgment' in a liberal democracy rests in part on the understanding that a decision is 'rendered by a fellow human being' and that it is seen as the product of 'a judicial decision

---

[17] John Morison and Adam Harkens, 'Re-engineering Justice? Robot Judges, Computerized Courts and (semi) Automated Legal Decision-Making' (2019) 39(4) *Legal Studies* 618, SSRN: <https://ssrn.com/abstract=3369530> accessed 7 April 2025 (citing to SSRN), 15.

[18] Ibid 9.

[19] Ibid.

[20] Ibid 19.

[21] Tania Sourdin and Richard Cornes, 'Do Judges Need to Be Human? The Implications of Technology for Responsive Judging' in T Sourdin and A Zariski (eds) *The Responsive Judge* (Springer 2018), <https://doi.org/10.1007/978-981-13-1023-2_4> 88; Cofone cites not this chapter, but Tania Sourdin, 'Judge v Robot? Artificial Intelligence and Judicial Decision-Making' (2018) 41(4) *UNSWLJ* 1114. I draw on Sourdin and Cornes for their more extensive exploration of similar points.

[22] Sourdin and Cornes ibid 88.

[23] Ibid 97.

[24] Ibid 99.

[25] Ibid 100.

maker struggling conscientiously, in public, with the detail of a case'.[26] A computer might possess the 'rational' ability to make a decision, but it would lack the 'emotional authority' to do it legitimately. Judgement operates under a wider set of cultural expectations. A fair decision is one made by a human who acts impartially, applying empathy and understanding to decide what justice calls for in a given case. No machine or program is likely to displace this assumption.

In a seminal contribution to this debate, Ian Kerr and Charissima Mathen asked whether a robot could perform the role of Chief Justice Roberts and whether such a machine could gain recognition as a legitimate replacement.[27] Legal reasoning, they contend, 'cannot be reduced to mere functional capabilities regarding extraordinary information gathering, speed, memory, recall, and even the ability to distinguish and disambiguate relevant legal rules.'[28] This is so for a number of reasons. One has to do with a distinction Wittgenstein drew between following a rule and complying with it. Kerr and Mathan contend that we understand human judgement or decision-making as entailing more than conduct that is consistent with a rule but also and primarily conduct that follows a rule. Following a rule involves feeling a moral compulsion, a 'normative weight,' to act in a certain way or reach a certain result, one that is derived from experience, social practice, or custom.[29] Paraphrasing Wittgenstein, they write: 'To follow rules is to adopt a particular *form of life*.'[30] Without feeling the greater normative weight of one principle over another that might be at stake in a given case, a machine would be incapable of deciding which should prevail and why.

Kerr and Mathen see a similarity between Wittgenstein's ideas about rule following and compliance and H.L.A. Hart's concepts of an internal and external point of view on the rules governing a society.[31] This points to a further limitation of AI in judgment. As Hart had argued, a person who adopts an external point of view on the rules can predict how people will

---

[26] Ibid 98, citing Micheal Kirby, 'The future of courts—do they have one?' (1999) 8 *J Judicial Adm* 383.

[27] Ian Kerr and Carissima Mathen, 'Chief Justice John Roberts is a Robot' (2014) University of Ottawa Working Paper, SSRN <https://ssrn.com/abstract=3395885> accessed 7 April 2025.

[28] Ibid 39.

[29] Ibid 25-26.

[30] Ibid 25, citing Ludwig Wittgenstein, *Philosophical Investigations* (GEM Anscombe trans, Prentice Hall 1973), §§ 241-242: 'It is what human beings *say* that is true and false; and they agree in the *language* they use. That is not agreement in opinions but in form of life.'

[31] Kerr and Mathen (n 27) 27-28, citing HLA Hart, *The Concept of Law* (OUP 1961) 87-88.

conduct themselves and when a 'hostile reaction or punishment' is likely to follow certain acts.[32] But to understand why people willingly choose to follow rules, why they embrace them as a 'standard of behavior and an obligation', one must adopt an internal point of view.[33] We do so when we consider the rules to reflect a normative framework, a set of propositions *worth* upholding. A person cannot play the role of a judge in a legal order without adopting this internal point of view, since deciding what *should* be done in a given case or how the law should apply in some novel situation, requires an understanding of — and commitment to — the normative assumptions that prevail in that society. A machine might be programmed to acknowledge these norms, but to choose one norm over another, one must *feel* a moral compulsion to do so; and this only a human, with internal point of view, can do.[34]

Finally, drawing on Dworkin, Kerr and Mathen contend that legal judgment 'requires being a *member of the community*, understanding its history, its moral convictions, having a point of view about its current character and having a stake in its future.'[35] Judgement requires this because it involves choosing in a given case the outcome that best fits with the web of a community's normative commitments and principles, and its sense of justice and fairness.[36] Yet even if a machine could do this, it could not share a judge's desire to decide a case in a certain way because as 'part of that community' she feels a personal stake in justice being done there.[37] Automated systems, robots, cannot conceivably do any of the things Wittgenstein, Hart, and Dworkin contemplate — and certainly not the forms of AI known to Kerr and Mathen at the time. They cannot follow a rule rather than merely comport with it, assume an internal point of view, or pass judgment as a member of a community with a holistic sense of its moral commitments and a compulsion to uphold them.[38] AI lacks the moral compass and drive to make decisions one way or another. In challenging cases, ones that cannot be resolved by direct application of a rule but only by crafting a new rule based on some larger principle, AI would only seem capable of deciding the case arbitrarily.

---

[32] Hart 87, cited in Kerr and Mathen (n 27) 27-28.

[33] Hart ibid 87-88, cited Kerr and Mathen (n 27) 28.

[34] Kerr and Mathen (n 27) 28.

[35] Ibid 39.

[36] Ibid at 30-39, discussing Ronald Dworkin, *Law's Empire* (Harvard University Press 1986).

[37] Kerr and Mathen (n 27) 37.

[38] Ibid 39.

Rebecca Crootof, writing in 2019, was also skeptical. 'Absent a breakthrough in machine intelligence [...an] idealized AI judge is infeasible.'[39] AI is limited to 'detecting patterns in data' and works well in 'narrow, limited settings, like chess' where there are 'clear right or wrong answers' or 'discernable underlying patterns and structures'.[40] It can calculate taxes or assess whether a contract has been breached but it cannot 'apply legal rules in accordance with changing social mores.'[41] The judgement we employ in common law courts is 'a distinctively human skill.'[42] It requires being 'sensitive to context, both to extenuating circumstances in individual cases and shifts in social norms over time,' and an ability to 'flexibly apply legal rules.'[43] While an AI system might be trained or programmed to incorporate context, 'that is hardly the same as having human contextualization at the time the algorithmic rule is applied, especially as that application may occur in a temporally, geographically, and culturally different context.'[44] Human judges can also be transparent about their reasoning, while AI judges 'may be black boxes'; and human judges 'internalize social norms' that inform their decisions and in turn strengthen those norms — and all of these human features are essential to judgment being perceived as legitimate.[45] A still further limitation is that while AI might apply an unambiguous rule to a given set of facts, it will founder when applying an ambiguous rule to a novel or complex situation.[46]

Standing back from this brief survey, the general thrust of the pre-2022 consensus on AI's limitations in judgment are clear: AI is good at pattern recognition, applying clear rules, generating probabilities; it is not good at normative judgement, novelty, social context, or

---

[39] Rebecca Crootof, ''Cyborg Justice' and the Risk of Technological-Legal Lock-In' (2019) 119 *Columbia L Rev Forum* 233 at 237.

[40] Ibid 237-8.

[41] Ibid 237.

[42] Ibid 238.

[43] Ibid.

[44] Ibid.

[45] Ibid.

[46] Ibid 239.

emotion. Cofone's survey cites other figures supporting this consensus, including Henry Surden,[47] Lisa Shay et al,[48] Cary Coglianese and David Lehr.[49]

Cofone's chapter also considers a consensus on two other notable concerns about AI's role in judgment: algorithmic bias and opacity. Both continue to be an issue in the era of generative AI. Yet Cofone's framing of them is worth canvassing to highlight their close connection to the state of the technology over time. In ways to be explored below, the focus of these concerns and possibilities for mitigation are evolving with AI itself. To understand how this is so, it is worth briefly looking back at the state of the question before 2022.

Drawing on his own research on point, Cofone distinguishes three forms of algorithmic bias that can result in 'inaccurate and discriminatory results'.[50] These include 'bias in the process of building the algorithmic model, bias in the sample that is used to train the algorithm, and societal biases captured and amplified by the algorithm.'[51] The first form can arise through the manner in which programmers choose to frame a problem or through decisions about what an algorithm should predict.[52] The second form of bias arises when AI is trained on a dataset that is 'unrepresentative of the population' or contains 'partial or incorrect data', resulting in 'problems at higher rates for an entire protected group compared to others'.[53] The third form

---

[47] Harry Surden, 'Artificial Intelligence and Law: An Overview' (2019) 35(4) *Ga St U L Rev* 1305, 1322-3: 'In short, current AI technology tends to work best for activities where there are underlying patterns, rules, definitive right answers, and semi-formal or formal structures that make up the process. By contrast, AI tends to work poorly, or not at all, in areas that are conceptual, abstract, value-laden, open-ended, policy or judgment-oriented; require common sense or intuition; involve persuasion or arbitrary conversation; or involve engagement with the meaning of real-world humanistic concepts, such as societal norms social constructs, or social institutions.'

[48] Lisa A Shay and others, 'Do Robots Dream of Electric Laws? An Experiment in the Law as Algorithm' in Ryan Calo, A. Michael Froomkin and Ian Kerr (eds), *Robot Law* (Edward Elgar 2016) 274, finding that 'Automation of legal reasoning is likely to be the most effective in contexts where legal conclusions are predictable because there is little room for choice in a given model; that is, they are determinable. Yet this experiment demonstrates that even relatively narrow and straightforward 'rules' can be problematically indeterminate in practice.'

[49] Cary Coglianese and David Lehr, 'Regulating by Robot: Administrative Decision Making in the Machine-Learning Era' (2017) 105 *Geo LJ* 1147, noting at 1177: '...even though algorithms can learn on their own and support automated decisions, humans still must decide how algorithms are specified, deployed, and integrated into broader administrative processes. [...] They do not set their own objective functions nor are they completely outside human control. An algorithm, by its very definition, must have its parameters and uses specified by humans'.

[50] Cofone, 'AI and Judicial Decision-Making' (n 9), drawing on Ignacio N Cofone, 'Algorithmic Discrimination is an Information Problem' (2019) 70(6) *Hastings LJ* 1389, 1399-1402.

[51] Cofone, 'AI and Judicial Decision-Making' (n 9) 8.

[52] Ibid 9.

[53] Ibid, also citing Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104(3) *Cal L Rev* 671, 680-681, 684-687.

of bias arises where data is representative of a population but 'produces a disparate impact outcome because of embedded social inequalities'.[54]

As Cofone notes, all three forms of algorithmic bias 'exist in practically every area of decision-making where AI is used'.[55] The best-known example, he points out, is the COMPAS software that parole boards in the United States use to assess a candidate's risk of reoffending. In a widely-cited investigative report, ProPublica revealed that COMPAS erred in identifying black offenders as high-risk at double the rate it erred with whites; and conversely, it was twice as likely to err in finding whites to be low-risk than blacks.[56] Further examples in the earlier literature of AI tools used in decision-making raising concerns about bias include software for credit scoring,[57] hiring and employment,[58] crime-mapping and policing,[59] and medicine.[60] Notably, all of them involve a limited form of AI that produces a predictive score or number, rather than the 'general purpose' form of AI involving language models. The three forms of concern about bias that Cofone outlines are at issue in the predictive score. The general concern found throughout the earlier literature he canvases is that the score that a piece of AI software might output will reflect a form of bias and there is no way to use the software without encountering these biases. They are a part of the algorithm at the core of the program. In Part II of this paper, I explore the question of whether the general purpose nature of language models enables users to mitigate bias to some degree through the way in which the models are used: what prompts are given and what material the model is asked to work on (but not what material it is trained on or how, which we know is subject to algorithmic bias).[61]

---

[54] Cofone, 'AI and Judicial Decision-Making' (n 9) at 9, also citing Aylin Caliskan, Joanna J Bryson and Arvind Narayanan, 'Semantics Derived Automatically from Language Corpora Contain Human-Like Biases' (2017) 356 (6334) *Science* 183; and Daniel Rosenberg, 'Data Before Fact' in Lisa Gitelman (ed), *'Raw Data' Is an Oxymoron* (MIT Press, 2013) 15.

[55] Cofone, 'AI and Judicial Decision-Making' (n 9) 9.

[56] Hulia Angwin and others, 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks' (*ProPublica*, May 2016) <www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> accessed 7 April 2025.

[57] Danielle Keats Citron and Frank Pasquale, 'The Scored Society: Due Process for Automated Predictions' (2014) 89(1) *Wash L Rev* 1.

[58] Pauline T Kim, 'Data-Driven Discrimination at Work' (2017) 58 *Wm and Mary L Rev* 857; Barocas and Selbst, 'Big Data's Disparate Impact' (n 53).

[59] Andrew D Selbst, 'Disparate Impact in Big Data Policing' (2017) 52 *GA L Rev* 109.

[60] David Lazer and others, 'The Parable of Google Flu: Traps in Big Data Analysis' (2014) 343 *Science* 1203.

[61] Emilio Ferrara, 'Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models' (2023) 28(11) *First Monday* <https://doi.org/10.5210/fm.v28i11.13346> accessed 7 April 2025, noting (at 1) biases

A further key concern with AI that Cofone identifies relates to opacity: a lack of transparency or knowledge of how an algorithm or AI software is formulated or designed to produce its output. As Cofone notes, scholars have distinguished three forms: 'intentional opacity, literacy-driven opacity, and inherent opacity.'[62] In the first case, companies are reluctant to disclose details about algorithms for competitive advantage; in the second, algorithms are too complicated for laypersons to understand; and in the third, 'a certain degree of opacity [...] is a natural and inevitable result of the characteristics of some algorithms.'[63] Once again, all of this applies to current forms of generative AI, giving rise to an already copious literature on point.[64]

The point to be emphasized here is that, as Cofone notes, for as long as AI has played a role in judgment and adjudication, algorithmic opacity has raised concerns about fairness.[65] The general idea hovering above these concerns is often unspoken in the literature on point. Algorithmic opacity conflicts with a broader expectation in liberal societies that the reasons for arriving at a decision be transparent, comprehensible, and justifiable in order to be *seen* as fair and impartial.[66] The principle is made more overt in the jurisprudence on the duty to give

---

'stemming from, among others, the nature of training data, model specifications, algorithmic constraints, product design, and policy decisions'; and Isabel O Gallegos and others, 'Bias and fairness in large language models: A survey' (2024) 50(3) *Computational Linguistics* 1097, exploring how 'models can learn, perpetuate, and amplify harmful social biases' (1097).

[62] Cofone, 'AI and Judicial Decision-Making' (n 9) 11, citing Jenna Burrell, 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) *Big Data Soc* 1, 1-2; Ignacio N Cofone and Katherine J Strandburg, 'Strategic Games and Algorithmic Secrecy' (2019) 64(4) *McGill LJ* 621, 626-632.

[63] Cofone, 'AI and Judicial Decision-Making' (n 9) at 11.

[64] An overview can be found in Manuel Carabantes, 'Why artificial intelligence is not transparent: a critical analysis of its three opacity layers' in Simon Lindgren (ed), *Handbook of Critical Studies of Artificial Intelligence* (Edward Elgar 2023) chapter 37; and Alberto Fernández, 'Opacity, Machine Learning and Explainable AI' in Francisco Lara and Jan Deckers (eds), *Ethics of Artificial Intelligence* (Springer 2024) 39.

[65] Cofone, 'AI and Judicial Decision-Making' (n 9) at 10, citing Margot E Kaminski, 'The Right to Explanation, Explained' (2019) 34(1) *BTJL* 189; Danielle Keats Citron and Frank Pasquale, 'The Scored Society: Due Process for Automated Predictions' (2014) 89(1) *Wash L Rev* 1; Rebecca Wexler, 'Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System' (2018) 70(5) *Stan L Rev* 1343; and Sonia K Katyal, 'Private Accountability in the Age of Artificial Intelligence' (2019) 66(1) *UCLA L Rev* 54.

[66] See, for example, Margot E Kaminski, 'Binary Governance' (2019) 92 *S Cal L Rev* 1529 canvassing the importance of justification in algorithmic regulation; Andrew D Selbst and Solon Barocas, 'The Intuitive Appeal of Explainable Machines' (2018) 87 *Fordham L Rev* 1085, discussing explanation of decisions as an inherent good; Reuben Binns, 'Algorithmic Accountability and Public Reason' (2018) 31 *Philosophy and Technology* 543, linking the normative case for algorithmic transparency in decision-making to the concept of public reason; and Kiel Brennan-Marquez, ''Plausible Cause': Explanatory Standards in the Age of Powerful Machines' (2017) 70 *Vand L Rev* 1249, noting at 1288: 'A key tenant of legality, separating lawful authority from ultra vires conduct, is the idea that not all explanations qualify as justifications.'

reasons.[67] The salient theme in the copious literature on algorithmic opacity — from Danielle Keats Citron's seminal 'Technological Due Process' in 2008,[68] to Frank Pasquale's *Black Box Society* in 2015,[69] to Cofone's more recent contributions on point[70] — is that while regulators and coders might take steps to reduce opacity, relying on algorithmic AI to make decisions will always be fraught with suspicions about hidden assumptions, biases, and framings buried deep in the code. As with bias, concerns about opacity persist with generative AI. But given that generative AI produces different forms of output by relying in part on a wider range of possible inputs (prompts, documents), it presents different possibilities for mitigation. In ways to be explored below, these will involve strategies more closely tailored to the new possibilities and challenges arising.

Before moving on to examine this argument in more detail, I close this part of the paper by noting (as Cofone did) that there were important dissenting voices among scholars of AI's role in law and judgment. Two that stand out are Ben Alarie and Eugene Volokh. Writing in 2016, Alarie made the case that given the inevitability that 'more observable phenomena' would become subject to 'greater quantification,' and that technologies of 'pattern recognition' would only continue to become more accurate, the two trends would culminate in what he called the 'legal singularity.'[71] This would be a state in which all legal uncertainty would be overcome, since law's application to an agreed set of facts would be immediately determinable

---

[67] For example, Justice Binnie, writing for the Supreme Court of Canada, in *R v Sheppard*, 2002 SCC 26, held [at para 15]: 'The courts frequently say that justice must not only be done but must be seen to be done, but critics respond that it is difficult to see how justice can be seen to be done if judges fail to articulate the reasons for their actions. Trial courts, where the essential findings of facts and drawing of inferences are done, can only be held properly to account if the reasons for their adjudication are transparent and accessible to the public and to the appellate courts.' See also HL Ho, 'The judicial duty to give reasons' (2000) 20 *Legal Studies* 42; *Coleman v. Dunlop Ltd*., [1998] PIQR 398 (Eng CA) 403; and *Flannery v. Halifax Estate Agencies Ltd.*, [2000] 1 All ER 373 (CA).

[68] Danielle Keats Citron, 'Technological Due Process' (2008) 85 *Wash U L Rev* 1249.

[69] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press, 2015).

[70] Ignacio N Cofone, 'Algorithmic Discrimination is an Information Problem' (2019) 70:6 *Hastings LJ* 1389 and Ignacio N Cofone and Warut Khern-am-nuai, 'The Overstated Cost of AI Fairness in Criminal Justice' (January 29, 2025; forthcoming Indiana Law Journal), SSRN: <https://ssrn.com/abstract=5117071> accessed 7 April 2025.

[71] Benjamin Alarie, 'The Path of the Law: Toward Legal Singularity' (May 27, 2016) 3, SSRN: <https://ssrn.com/abstract=2767835> accessed 7 April 2025.

or 'universally accessible in real-time.'[72] There may continue to be, as Alarie notes, 'disputes over facts, but once found, the facts will map on to clear legal consequences.'[73]

Alarie's conception of AI's possible impact on law was based in large part on the kind of prediction software he had a role in developing — tools that would predict whether a given provision of tax law would apply to a set of facts, or whether one would be found to be an employee or a contractor.[74] The model for automated law here takes a basic conditional form: if this, then that. But what happens if we cannot agree on the facts? What if the essence of a dispute is what normative significance to assign to a given fact? Anticipating this objection, Alarie speculated that it was at least possible that AI will eventually be able to produce a 'menu' of policy options or normative interpretations when formulating rules for new factual situations, and be able to choose the 'best' option among them in light of a given 'normative objective' supplied by a human.[75] In Alarie's vision, AI can be a powerful tool to assist humans in passing judgment, but not capable of performing all of their functions in judgment.[76]

In a remarkably prescient contribution from 2019, Eugene Volokh went further by envisioning a situation very close to the one we have now arrived at.[77] Volokh argued that at some point, AI may be able to produce a litigation brief at least as well as the average lawyer — based on persuasiveness as the main criterion of evaluation.[78] And if AI can produce a brief that is at least as persuasive as a brief done by a lawyer of average competence, and do so consistently, then it would at some point 'make sense to let the AI make the decision itself'.[79] An AI brief-writer would be technically competent to be an AI judge since it could 'construct persuasive arguments that support the various possible results in the case,' and 'choose from

---

[72] Ibid.

[73] Ibid.

[74] Alarie's involvement in developing an AI software company, Bluejay Legal, is recounted in the Introduction to Abdi Aidid and Benjamin Alarie, *The Legal Singularity: How Artificial Intelligence Can Make Law Radically Better* (University of Toronto Press 2023).

[75] Alarie 'Path of Law' (n 71) 10. Alarie would develop his theory, together with Abdi Aidid, in *The Legal Singularity* (n 74). The manuscript appears to have been completed prior to the appearance of ChatGPT in the fall of 2022, given that the book makes no reference to it, except in a two-page Afterward in which the authors describe OpenAI's new platform as having 'surpassed even our most optimistic predictions' for AI (203).

[76] Alarie, 'Path of the Law' (n 71) 10.

[77] Eugene Volokh, 'Chief Justice Robots' (2019) 68 *Duke Law Journal* 1135.

[78] Ibid 1141. I canvas in the following section generative AI's present capability to do precisely this.

[79] Ibid 1142.

all those arguments the one that is most persuasive, and thus the result that can be most persuasively supported.'[80] All of this could be tested using a panel of human judges to evaluate the performance; but in theory, nothing precludes AI from reaching the point at which it satisfies human judges about reaching the right result consistently, thus warranting an ongoing delegation of judgment.[81]

Volokh's argument rests on two assumptions. The first is that we set aside the concern about whether AI understands what it is processing and accept that the criteria for deciding whether AI should replace a human judge is whether it can produce a sufficiently persuasive decision. Volokh was overt on this point.[82] The second assumption underlying his argument is that a persuasive but purely automated decision on a matter of social or moral import could come to be accepted as legitimate in our current socio-political framework. He addressed this assumption implicitly by addressing various concerns that arise in response to it. One is to suggest that parties with reservations about AI taking over the adjudicative role of a human judge might be given the chance to opt out; *i.e.*, no AI decision-making without both parties consenting.[83] Another is to suggest that if legitimacy in judgment depends on perceptions of a decision reflecting 'compassion, wisdom, mercy, [or] justice', the AI could be programmed to do precisely this.[84] It would remain to be seen whether a sufficiently compassionate judgement produced by AI would prove to be acceptable, but Volokh was optimistic that it could be.[85]

Volokh's optimism about AI judges extended to other concerns. He argued that racial or other bias derived from training data or resulting through 'emergent properties' of AI could be overcome by programming and design.[86] Whether modifications to code have resulted in AI avoiding bias is something that might be tested to confirm a degree of impartiality, at least

---

[80] Ibid 1141.

[81] Ibid 1142.

[82] Volokh takes as his point of departure the sufficiency of technical competence as a qualification for playing a given role: 'If an entity performs medical diagnoses reliably enough, it's intelligent enough to be a good diagnostician, whether it is a human being or a computer. We might call it 'intelligent,' or we might not. But, one way or the other, we should use it. Likewise, if an entity writes judicial opinions well enough [...] it's intelligent enough to be a good AI judge' (1138).

[83] Ibid 1160.

[84] Ibid 1167.

[85] Ibid 1170-71.

[86] Ibid 1168, asserting 'The AI judge might, for instance, be programmed to ignore certain attributes, such as parties' race, in drawing its generalizations. The training data might also be vetted to minimize bias flowing from that data'.

comparable to human judges.[87] And while some have argued that only human decisions will be perceived to be legitimate based on a belief that dignity requires being heard by a human, the opposite might also be true.[88] In some cases, an AI judge might seem more credible or legitimate on the basis of complete disinterest or impartiality — not being friends with one party's counsel, not being concerned about reelection, or not being biased on the basis of race or gender.[89] AI's reasons would be so detailed and transparent as to instill confidence and respect, and after the initial novelty of AI judgement wore off, the public would largely come to accept automation in this role as legitimate and normal.[90]

II. Generative AI in judgment and mediation

Since being made available to the public in the fall of 2022, forms of text-based generative AI have begun to be used in the practice of law in countless ways, including research, writing, and file management. Some of these uses involve capabilities anticipated in the earlier literature on AI in judgment, and some do not.[91] Judges have been more trepidatious about embracing AI. A report in 2023 about a Columbian judge using ChatGPT to decide a case gained notoriety, but remains exceptional.[92] Government agencies and administrative tribunals in nations around the world are, however, making various uses of AI to help sort files, speed up processes, and

---

[87] Ibid 1169.

[88] Ibid 1170. Volokh does not cite any dignitarian arguments for a human judge, but one example can be found in Mimi Zou and Ellen Lefley, 'Generative Artificial Intelligence and Article 6 of the European Convention on Human Rights: The Right to a Human Judge?' (2024) UNSW Law Research Paper No. 25-8, SSRN: <https://ssrn.com/abstract=5040351> accessed 7 April 2025, arguing at 3 that 'Given courts are public institutions and the ultimate authority in resolving civil disputes, we find that a dignity-based interpretation of Article 6 [of the ECHR] includes the right to a human judge.'

[89] Volokh (n 77) 1169.

[90] Ibid 1170-71.

[91] Tools such as Google's NotebookLM, or other language models which can work on documents one uploads, can perform myriad tasks in relation to them (summary, compression, revision, creating podcasts, etc.) — a suite of capabilities seldom envisioned or contemplated in the earlier literature, possibly because it would have been hard to imagine attaining all of this in relatively short order.

[92] As one account notes, the judge posed questions to Chat GPT, including the central issue in the case: 'Is an autistic minor exonerated [in Columbian law] from paying fees for their therapies?' Luke Taylor, 'Colombian judge says he used ChatGPT in ruling' (*The Guardian* February 2023) <https://www.theguardian.com/technology/2023/feb/03/colombia-judge-chatgpt-ruling> accessed 7 April 2025. Courts elsewhere have conducted various experiments using chatbots in judgment, including in Pakistan, China, and the United States: André Guskow Cardoso and others, 'Generative Artificial Intelligence and Legal Decision making' (2024) 19(11) *Global Trade and Customs Journal* 710, 711, noting Judge Newsome's use of ChatGPT to help define the meaning of a word in *Snell v. United Specialty Insurance Company* (22 12581, USCA11) (28 May 2024).

assist in adjudication — though in most cases not using AI to make critical decisions.[93] All of this is background.

In the first segment of this Part of the paper, I single out an experiment conducted by Adam Unikowsky in mid-2024.[94] Unikowski sought to demonstrate that generative AI could do far more than assist a judge with facets of adjudication, such as research. It can now draft opinions and decide appellate cases to a strikingly effective degree. Reading his experiment in light of the earlier literature canvased in Part I shows how AI's role in adjudication can vastly exceed predicting outcomes or probabilities. It offers evidence that AI can produce the outline of an opinion in a given case that contains reasons for judgement that approach the quality and sophistication of those found in decisions rendered by an apex court. This includes an ability to formulate new legal tests and apply them to the facts in the case or to hypothetical situations. The experiment also shows that AI can be prompted to be more creative in fashioning rules, tests, and reasons than those proposed by the parties in the case or found in prior case law.

In what follows, I briefly canvas Unikowsky's experiment before outlining my own efforts to replicate and expand on portions of it in the Canadian context. I then discuss how the results challenge assumptions in the literature canvased in Part I of this paper and conclude with comments about remaining limits and reservations about AI in judgment.

a. Experiments in appellate judgment

Unikowsky set out to demonstrate AI's capabilities in appellate judgment using Anthropic's Claude Opus 3 model, its most advanced model at the time.[95] His method was to upload to Claude briefs submitted in recent cases before the United States Supreme Court in the current term and to ask Claude to outline a decision in three to four paragraphs.[96] In the interests of space, I have chosen not to reproduce Unikowsky's comparison of Claude's summary opinions (which he includes in full) with the Court's decisions. I note, however, that his write-up contains Claude's summary opinion in six cases the Court had decided the week prior, which

---

[93] Guskow Cardoso and others (n 92).

[94] The experiments are detailed in Adam Unikowsky, 'In AI we trust', June 8, 2024, <https://adamunikowsky.substack.com/p/in-ai-we-trust> accessed 7 April 2025 and 'In AI we trust, part II', June 16, 2024, <https://adamunikowsky.substack.com/p/in-ai-we-trust> accessed 7 April 2025. I became aware of these posts in Guskow Cardoso and others (n 92) 711-12.

[95] Unikowsky, 'trust part II' (n 94).

[96] As his posts indicate, in some cases, Unikowsky uploaded only the two main briefs and in others he added three or four additional briefs.

provide a good example of the degree of sophistication in legal reasoning it is capable of replicating on a range of issues. In Unikowsky's assessment, 'Claude 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.''[97] Among the 37 merit cases the Court had decided to that point in the term, Unikowsky found that Claude decided 27 of them the same way the Court did.[98] In the remaining 10, Unikowsky writes that he 'frequently was more persuaded by Claude's analysis than the Supreme Court's'.[99]

Anticipating the objection that Claude was merely summarizing the briefs and could not come up with creative solutions to legal disputes, Unikowsky focused on what Claude did in a case calling for the creation of a new legal test. Briefly, in *Lindke v. Freed*,[100] the Court was tasked with deciding whether government officials who block followers on social media violate the First Amendment. Did a city manager who deleted comments on his Facebook page engage in state action? Unikowsky prompted Claude to resolve the case by formulating 'the appropriate legal standard to be applied to disputes of this nature.'[101] Claude produced an opinion summarizing the facts in one paragraph, the Court's earlier jurisprudence on point in another, and then set out a test in the third and final paragraph:

> The central question is whether the public official has used the social media account as an extension of his office – directing governmental functions through the account and operating it in a manner that conveys it is an official government presence. Factors to consider include whether the account is styled in a manner that closely identifies it with the official's governmental position, whether the account has been used as a tool for communicating official policy and statements, whether government resources are used to maintain the account, and whether the account is used as a platform for interacting with constituents on matters related to the officials governmental duties.[102]

Unikovsky notes the similarity to the test the Court actually formulated in this case:

---

[97] Unikowsky, 'trust part II' (n 94).

[98] Ibid.

[99] Ibid.

[100] *Lindke v Freed*, 601 US 187 (2024).

[101] Unikowsky, 'trust part II' (n 94).

[102] Ibid.

> A public official who prevents someone from commenting on the official's social-media page engages in state action under §1983 only if the official both (1) possessed actual authority to speak on the State's behalf on a particular matter, and (2) purported to exercise that authority when speaking in the relevant social-media posts.[103]

Missing here were the factors that Claude included, but Claude's opening three lines (setting out the main test) captures the crux of the Supreme Court's test: conveying the appearance of, or purporting to exercise, speech in an official capacity.

Unikowsky goes on to demonstrate Claude's further capabilities in judgment by asking it to consider a hypothetical raised in the Supreme Court decision itself — a facet of the experiment relevant to claims about AI's reasoning limits canvassed in the earlier literature. The hypothetical involves a school board president who announces at a board meeting the lifting of pandemic-era restrictions on schools, and then shares this with neighbours the next evening at a barbecue. Tasked with applying the legal standard it had created moments before, but without relying on uploaded briefs, Claude could readily apply its test to this scenario and provide several factors supporting why the statement at the board meeting was state action but not what was said at the barbecue. Claude was then asked to compare the legal test it came up with to the one in the Supreme Court's decision, and to explain 'which standard is better and why?' Claude offered three substantive differences in the way each operates and four reasons its own test is superior — persuading Unikovsky: 'This is an excellent answer — very thoughtful and directly responsive to the question.'[104] To go a step further, Unikovsky asked Claude to 'offer a creative, alternative legal standard that neither party proposed that might have some merit', followed by new arguments for and against it. The test Claude came up with:

> A public official's social media activity should be considered state action if either (1) the activity occurs on an account that is government owned or controlled and used for official purposes, or (2) The activity occurs on a personal account that has been used persistently and pervasively for official purposes, such that the account has become 'governmental in nature.'[105]

Claude followed this with four substantive reasons in favour of adopting the standard and four against, followed by a conclusion that pointed to a middle ground: 'Some refinement to lower

---

[103] Ibid, citing *Lindke* (n 100) 187.

[104] Unikowsky, 'trust part II' (n 94).

[105] Ibid.

the bar for personal accounts and incorporate the viewer's experiences may help thread the needle.'[106] As Unikowsky observed: 'With no priming whatsoever, Claude is proposing a completely novel legal standard that is clearer and more administrable than anything proposed by the parties or the Court.'[107]

I sought to replicate Unikowsky's experiment using one of OpenAI's most advanced language models in early 2025, GPT-4.5 involving two decisions at the Supreme Court of Canada. One of them has been decided, the other is pending.

The first is *R v Bykovets*,[108] a case decided in late 2024 that addressed the question of whether a person has a reasonable expectation of privacy in an IP address. This was a 5-4 decision finding that an IP address does engage a privacy interest. It was a good candidate for the experiment because it could have gone either way — an important fact that points to remaining potential limits of AI in judgement canvassed further below.

I uploaded seven of the factums in the case (the parties and various intervenors) and asked the model: '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 a person has a reasonable expectation of privacy in an IP address.' GPT-4.5 produced a text headed 'Outline of the Decision of the Court' and divided into four headings: 'Introduction and Overview of the Appeal,' containing an apt and concise factual summary; 'Analysis of the Legal Framework,' which set out correctly the main test for deciding whether a person has an expectation of privacy in something; 'Consideration of Competing Arguments,' which concisely and correctly summarized the position in all 7 factums; and the 'Court's Decision and Reasoning.' This paragraph reads:

> The Court finds that an IP address, on its own, does not attract a reasonable expectation of privacy. While it is true that an IP address is a necessary component of online communication, it does not inherently disclose private information unless combined with other data. The police in this case did not obtain subscriber information without a production order, in compliance with Spencer. The Court agrees with the majority of the Alberta Court of Appeal that an IP address alone is not part of an individual's biographical core information and, therefore, does not warrant constitutional

---

[106] Ibid.

[107] Ibid.

[108] *R v Byokovets*, 2024 SCC 6.

> protection. The Court is mindful of the need to balance privacy interests with law enforcement imperatives and concludes that requiring judicial authorization to obtain IP addresses would create an undue burden on police investigations without significantly enhancing privacy protections.

This is in essence precisely what the dissent held in *Bykovets*. The two core holdings in Justice Côté's dissenting opinion were: '...an IP address, on its own, reveals only limited information. It does not reveal a 'biographical core of personal information'...'[109] and '[t]he result reached by [the majority] that not only some, but all, IP addresses attract a reasonable expectation of privacy … would seriously thwart the police's ability to investigate such serious offences against children.'[110]

But, as noted, *Bykovets* was a 5-4 decision and happened to go the other way. I prompted the GPT-4.5: 'Outline a decision that goes the other way.' Here again, the model's summary decision closely tracked the Court's majority opinion. Divided into similar headings, the crucial segment stated:

> The Court concludes that individuals have a reasonable expectation of privacy in their IP addresses for the following reasons:
>
> 1. *An IP address can reveal private and biographical information* – While an IP address alone may not identify a person, it enables the state to track and analyze a user's online behavior, raising significant privacy concerns.
> 2. *Online anonymity is a protected interest* – The ability to browse the internet without fear of constant surveillance is fundamental to the right to informational privacy.
> 3. *Judicial oversight is necessary* – Given the potential for misuse and the growing sophistication of data analysis techniques, law enforcement should be required to obtain prior judicial authorization before acquiring IP addresses from third parties.
> 4. *Section 8 protections must evolve with technology* – In an era of mass data collection, privacy protections must be interpreted broadly to ensure that individuals retain control over their personal information.

---

[109] Ibid [129].

[110] Ibid [160].

The propositions contained in each of these four points can be found in the headnote of the Court's actual decision summarizing the majority opinion:

1. *IP addresses reveal biographical info:* '...the purpose of s. 8, appreciated normatively, requires asking what information the subject matter of the search tends to reveal. As the link that connects specific Internet activity to a specific location, an IP address may betray deeply personal information, even before police try to link the address to the user's identity.'
2. *Online anonymity is a protected interest:* '...An IP address can also set the state on a trail of Internet activity that leads directly to a user's identity, even without a warrant compelling an ISP to disclose the name and address of the customer for the IP address.'
3. *Judicial oversight is necessary:* 'Weighed against these substantial privacy concerns is society's sometimes conflicting but legitimate interest in the need for safety and security. Police should have the investigative tools to deal with crime that is committed and facilitated online. However, requiring that police obtain prior judicial authorization before obtaining an IP address is not an onerous investigative step.'
4. *Section 8 protections must evolve with technology:* 'A court must take a holistic view of the subject matter and must be especially careful in describing the subject matter of a search touching electronic data. The approach must not be mechanical, and it must reflect technological reality.'

Before moving on to observations about this, I note one reservation. Despite my having prompted GPT-4.5 to draw only on the factums to produce its summary opinions here, it is not clear that the model has not drawn on the Court's actual decision through the incorporation of that text in its training data.[111] The only way to test the quality of output in the absence of a real decision would be to ask a language model to produce a summary decision in a case in which factums have been provided but the Court has not yet rendered its decision.

At the time of this writing, the Supreme Court of Canada has heard arguments in *R v Singer*,[112] but not rendered its decision. This is a case about whether police conducted an

---

[111] Evidence of the technical possibility of this is canvassed in A. Feder Cooper and James Grimmelman, 'The Files are in the Computer: On Copyright, Memorization, and Generative AI' forthcoming 2025 *Chicago Law Review* (arXiv:2404.12590), arguing that in the process of training, language models do not merely glean statistical patterns among words in a text, but 'memorize' or copy and store entire texts — as demonstrated through the processes of 'extraction' and 'regurgitation', where a model produces a complete or near-complete copy of a text.

[112] *R v Singer*, SCC case 41090. The Court heard arguments in February of 2025; I conducted my test in March, 2025.

unlawful search when they came onto a private driveway to investigate a report about an impaired driver and found the driver asleep in their vehicle. I uploaded to GPT-4.5 nine of the factums in the case and asked: '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.'[113] The model produced an 'Outline of the Court's Decision' in a similar format to the one it provided for *Bykovets*. The key portions were those pertaining to whether Singer had a privacy interest while in his vehicle on his driveway. GPT-4.5 found that:

> The implied licence doctrine, which allows certain intrusions onto private property, does not apply because the police approached expressly to investigate a criminal offense without invitation or exigent circumstances. Accordingly, the police conduct constituted a 'search' within section 8, and due to the lack of warrant or clear justification, it was unreasonable.

The model also considered but rejected a line of argument urged by the prosecution:

> The Court rejects the argument that the reduced expectation of privacy in vehicles universally justifies warrantless searches. Given that the vehicle was on a private driveway, the expectation of privacy was closer to that of a private dwelling. The police could have pursued less intrusive methods, and the circumstances did not present urgency sufficient to justify the warrantless entry.

Finally, applying the test for exclusion of evidence under section 24(2) of Canada's Charter, GPT 4.5 concluded:

> …the breach was serious, involving deliberate intrusion onto private property without lawful authority. The impact on Mr. Singer's Charter-protected interests was significant, infringing upon the privacy afforded to private residences and their surroundings. While society has an interest in adjudicating impaired driving cases, it also has a vested interest in police accountability and protecting individual liberties. On balance, admitting evidence obtained through a clear and deliberate Charter violation would

[113] I recognize that running this experiment before the Supreme Court of Canada has decided *Singer* does not entirely address the concern about drawing on a memory of a decision in the data, since GPT 4.5 might have memorized (might contain a copy or substantial portions of) the Court of Appeal decision in this case — a point I address below.

> undermine public confidence in the judicial system. Thus, the evidence must be excluded.

Following this query, I prompted GPT 4.5 with the same question I asked in relation to *Bykovets*: 'Now decide the case the other way.' It produced an outline in a similar format and length. It found that the implied license doctrine did apply and that public safety concerns outweigh the privacy interest here, rendering the search reasonable. It also held that even if a minor breach had occurred, it would not warrant exclusion under 24(2). I suspect that when the Court's actual decision is rendered, the two outlines — the first followed by the second — will map closely on to the gist of the majority and dissenting opinions, respectively.

But the majority or dissent might do something different. It might fashion a new police power. There is a basis for this in the factums filed by the Attorneys General of Saskatchewan, Ontario, and British Columbia, but only one factum (Ontario's) proposed a precise test to be formulated. Asking GPT-4.5 to outline a test for this is analogous to the portion of Unikowsky's experiment in which he asked Claude to develop a new and more creative test for when a public figure in a private setting engages in state action. 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.

Its proposed test was impressively creative and novel. The main paragraph stated:

> 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.

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. The Court of Appeal decision in *Singer* also

only treats the ancillary powers document briefly, dismissing the argument that it should apply here.[114] And without being explicitly prompted to do so, GPT-4.5 went a step further and applied the 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 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.

In the interests of space, I have not proceeded to ask GPT-4.5 to provide reasons that would justify *recognizing* a new ancillary police power here, but this would be part of any full decision in which a court decided to go this route. However, the test outlined here, along with Unikowsky's original experiment, demonstrates that a language model is capable of formulating a new multi-part test, providing reasons to justify it, and applying it to a given set of facts. It can also do so at a level of sophistication that approximates the reasoning found in real decisions of an apex court.

b. Observations and implications

What, then, does this experiment reveal about generative AI's capability in legal judgment more broadly? Does it unsettle or challenge any of the assumptions about AI's limitations canvassed in Part I of this paper?

A first observation is that Unikowsky's experiment and my efforts to replicate it only involved the creation of an *outline* of an apex court decision, not the draft of a full decision itself. A full opinion would contain a nuanced and textured presentation of detail and argument. This would include a more detailed procedural history of the case (trial and appeal decisions below) and the decades-long body of case law on point leading up to the present appeal. A full opinion would also include the more elaborate analysis and justification that an appellate court offers for making the ruling on crucial issues that it does — rather than a mere paragraph summarizing this. Can generative AI do all of this?

---

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

There are two reasons to infer that, at this point, it can or soon will. First, as Unikowsky showed with Claude 3 in mid-2024 and as I have demonstrated above with GPT-4.5 in early 2025, frontier language models are good enough now to be capable of reading and analyzing party material and producing a draft decision that meets one specification or another. I followed Unikowsky in prompting a language model to produce only an outline of a decision and asking it to generate a different legal test (in my case, on ancillary powers) from the one proposed by parties in the case. This offered evidence of a language model's ability to perform the legal analysis — or its facsimile — that forms the core of a legal judgment. But neither of us tested a model's ability to draft longer reasons, in a style or format approximating a full decision. Claude or GPT-4.5 might be capable of producing a draft of one; but a better tool for this is on the horizon, pointing to a second reason to be optimistic about AI's capability here.

New tools such as OpenAI's Deep Research that have emerged in recent months can produce detailed, twenty or thirty-page research reports structured not unlike an appellate court decision — with a summary outline, followed by sections containing a survey of content on point, followed by analysis and conclusions.[115] A key feature of these tools is that they draw on dozens of external sources and contain footnotes to those sources. At the moment, these are limited to material found on the open web, but OpenAI has indicated an intention to enable Deep Research to cite documents found in proprietary or specialized databases.[116] If and when Deep Research or similar tools are connected to a case law database and provided the procedural history of a case, they can be prompted to produce a lengthy and detailed decision that incorporates all of this material, along with additional case law or passages from cases not cited in party materials. Draft opinions from these more advanced 'chain of reasoning' models could also include a more fulsome or nuanced analysis and justification for holdings on the central issues. The quality of the output of OpenAI's Deep Research at this point in time points in the direction of there being no *technical* impediment to producing a judgment approximating a first draft of a full apex or appellate court decision rather than a brief outline.[117]

---

[115] Alex Olteanu, 'OpenAI's Deep Research: A Guide With Practical Examples', 5 February 2025, Datacamp, <https://www.datacamp.com/blog/deep-research-openai>.

[116] OpenAI, 'Deep Research FAQ' (OpenAI February 2025) <https://help.openai.com/en/articles/10500283-deep-research-faq> accessed 7 April 2025.

[117] Until Deep Research and tools of this kind are connected to legal and other third-party databases (rather than merely drawing on sources found on the open web), an experiment involving Deep Research writing lengthy court

But what to make of the capability for legal reasoning and judgment that generative AI has demonstrated in the Unikowsky experiment summarized and replicated here? Unikowsky's own conclusions mirrored Eugene Volokh's optimistic predictions of AI's adequacy and suitability for judgment canvased earlier.[118] Appellate judging, as Unikowsky notes, essentially entails taking 'text inputs,' applying 'logical reasoning' and producing 'text outputs.'[119] In most cases, this does not involve elaborate or complex forms of reasoning; issues at the core of a case are often simple or straightforward. It is only when a case raises a difficult moral or normative question on which reasonable people might disagree that an AI decision might seem inappropriate. Unikovsky's response to this conundrum is to argue that 'it's impossible for AI to be worse than human judges, because no matter how the AI decides the case, it will match the output of at least some human judges.'[120] Volokh's position was close to this. He argued that if we measure the quality and effectiveness of an AI decision based on the persuasiveness of its reasons, it would likely be capable of doing no worse than the average human judge in taking a position on a moral issue, since he or she can do no more than offer cogent and persuasive reasons in defence of it.[121] AI, in short, can at least map out or even draft an entire opinion in accordance with a moral position determined by a human judge — a capability and role that Benjamin Alarie, among others, had anticipated.[122]

But would AI in this judgment-drafting role be capable of mitigating other concerns, including doubts about the ability to reason by analogy, to follow rather than conform with a rule, to be sensitive to context (such as rapidly changing norms), and to avoid bias and opacity in judgment? The experiment canvassed here suggests that language models can indeed mitigate all of these concerns to a significant degree, though not completely. They do so in two ways: by acting primarily on materials submitted by parties before the court and by providing cogent and persuasive reasons for judgement. The experiment also demonstrates that, given its

---

judgments is premature. But carrying out this experiment is an obvious next step for research in this area, with courts and tribunals around the world likely to be at the forefront of such experimentation as well.

[118] Unikowsky, 'In AI we trust' (n 94); Volokh (n 77).

[119] Unikowsky, 'In AI we trust' (n 94).

[120] Ibid.

[121] Volokh (n 77), asking at 1154: 'What more can we reasonably ask of an opinion drafter—human or AI—than the production of opinions that a blue-ribbon panel of trained observers will accept over the alternatives?'

[122] Alarie (n71) 10.

greater facility with language and forms of reasoning, AI could take over more facets of judgment than was contemplated in the earlier literature.

To canvas the first concern noted in Part I about AI's role in judgment, recall Sunstein's argument about AI's limitations in legal reasoning.[123] He argued that AI can assess a degree of similarity between cases based on proximity in terms of certain variables — indicating, for example, how similar one case about national security is to earlier cases in that category, or the probability of an outcome based on a degree of similarity. But it cannot decide whether the *category* of national security cases — over, say, free speech cases — is the right class, or contains the right set of variables, for comparison. This involves a value judgment: *i.e.*, deciding that the present case is *really about* national security rather than free speech, that its facts are better understood as an instance of the larger principle that courts should defer to the government when a threat at issue is significant. It involves deciding that this case is more so an instance of one principle than other principles that might apply. Sunstein assumed that this kind of decision had to be made by a human at the outset. Can generative AI do this?

An example of AI engaging in this kind of reasoning in the experiment conducted above can be found in the discussion of *Bykovets*. Asking the Supreme Court and GPT-4.5 to decide whether we have a privacy interest in an IP address entails asking: Is a police demand for an IP address more like police asking a gym owner to open a person's locker? Or is it more like police visiting the gym and asking to see the sign-up sheet for the squash court? Is it, in essence, an instance of the principle that police act unlawfully when they demand to see something private acting without authority, or it is more like cases where they ask to see something that isn't private and thus act lawfully? How would the reasonable person think of an IP address? How *should* they think of it in a free society? The experiments involving the *Bykovets* and *Singer* cases outlined above offered evidence that language models are capable of reasoning analogously in this way. Models can output an opinion finding that the facts in a case at issue are *better understood* as an instance of one general principle over another, and they can provide reasons to justify this.

Unikowsky's original experiment offers further evidence to support AI's capability in this regard. As noted, he found that Claude arrived at the same outcome in 27 of 37 US Supreme Court cases in the current term. A number of these cases involved an application of a larger principle to the facts in a given case, as Unikowsky's survey of the results in 6 of them

---

[123] Cass Sunstein, 'Of Artificial Intelligence and Legal Reasoning' (2001) 8 *U Chicago L Sch Roundtable* 29.

demonstrate.[124] An important caveat, however, is that in close cases — ones that resulted in 5-4 decisions among humans divided as to the right principle to apply, as in *Bykovets* — the model appears to choose arbitrarily. Why did GPT-4.5's first answer to being prompted to decide *Bykovets* result in a close match to what turned out to be the dissenting opinion in the case? Why did it not match the majority opinion first? Does it matter that we do not understand *how* AI makes this decision, even though it offers cogent and persuasive reasons for making that decision?

The answer depends in part on how we measure or assess arbitrariness; but however we do, important concerns are raised by *how* AI arrives at a legal decision. Under the hood, as it were, language models may decide a case one way or another based on the weight of opinion among factums on which they draw (how many side one way or another), or the amount of text devoted to a principle in those factums, or by drawing on reasons from cases the model may have memorized in its training data.[125] Ascertaining this would call for a separate study canvassing these ratios in all the briefs and factums of the various experiments outlined here and perhaps a larger sample involving a larger body of cases. Yet without more clarity on this, how a language model makes a legal decision — *how* not why — seems mysterious and opaque. Can we trust AI as a decision-maker, can we even assume it is *capable* of making legal decisions, if we cannot be sure about how they are made — even though, once again, they may be, on their face, carefully reasoned?

One response to this concern is to conclude that without knowing how, technically, AI arrives at its decisions, it cannot be trusted. Claude's performance in Unikowsky's experiment,

---

[124] *United States Trustee v John Q. Hammons Fall 2006, LLC*, 602 US ___ 2024 (applying the 'touchstone principle that congressional intent governs the remedial inquiry'); *Campos-Chaves v Garland,* 602 US ___ 2024 (noncitizens should be ordered removed in absentia after they fail to appear at their removal proceedings); *Garland v Cargill*, 602 US 406 (2024) (an item is a machinegun when it shoots more than one bullet per pull of the trigger); *FDA v Alliance for Hippocratic Medicine*, 602 US 367 (2024) (a plaintiff must establish standing by showing a particularized injury traceable to the defendant's actions and one likely to be redressed by a favorable decision); *Starbucks v McKinney*, 602 US ___ 2024 (injunctive relief should be granted in the labour relations context when it meets four conditions); and *Vidal v Elster*, 602 US ___ 2024 (a limit in trademark legislation is a justified infringement of expression if government purposes are valid and restriction is limited).

[125] The basis of this suspicion arises from research on biases that can shape the output of a language model. See for example, A Acerbi and JM Stubbersfield, 'Large language models show human-like content biases in transmission chain experiments' (2023) 44 *Proc Natl Acad Sci USA* 120, <https://doi.org/10.1073/pnas.2313790120>; and Emily M Bender and others, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), Association for Computing Machinery, New York, <https://doi.org/10.1145/3442188.3445922>, noting at 617: 'If the [language model] or word embeddings derived from it are used as components in a text classification system, these biases can lead to allocational and/or reputational harms, as biases in the representations affect system decisions.'

27 out of 37 cases decided the same way as the Court itself, is impressive but not reliable enough. Unless one could be certain that the AI cannot be gamed, that an outcome cannot be engineered by hacking the model in some way, it cannot be relied upon; and if so, it cannot be said to be *capable* of legal judgment. This is so because judgment on a legal matter requires not just technical competence or the ability to provide a reasoned opinion. It also requires confidence in the impartiality of the decision-maker. If AI's process can be hijacked by a technically savvier party, its neutrality cannot be assured. Until we gain further clarity as to *how* language models decide cases, this may be the strongest argument in favour of humans retaining supervision or oversight of AI in decision-making, of not — and perhaps never — relying on AI alone. Yet even so, this would only mean that AI is not capable or fully reliable of making legal decisions *alone*. It would not prove that AI could not at least produce a viable and close to complete draft of them.

Another response to this concern is to suggest that what matters is not how AI decides a case but what reasons it offers for why it does so. We should be concerned not with what is happening under the hood with AI, but with what an AI draft decision contains on its face. Eugene Vololk had made this point in 2019, invoking Turning's analogous argument in 1950.[126] We should not be concerned with whether AI understands what it is deciding but with whether its reasons for decision are cogent and persuasive.[127] Put another way, it should not matter whether AI is arriving at its decision because five out of six factums or briefs argued for one outcome over another, or whether one principle got more coverage in those materials than another. What matters in terms of how acceptable, legitimate, and transparent the decision will be said to be is how persuasive its reasons are to a given audience. Whatever is happening under the hood with Claude or GPT-4.5 in the experiments canvassed here, however readily it might be gamed (with knowledge of its inner workings), nothing rules out approaching the output of AI from Volokh's perspective. On this view, the experiment surveyed above demonstrates that AI is capable of producing the outline of a decision cogent and persuasive enough to appear justified, impartial, and transparent. And that is all that should matter for assessing AI's competence and capability in judgment. This may not settle the question of whether AI should be deciding cases entirely on its own; but it would settle the question of *capability*.

---

[126] Volokh (n 77) 1137-8, citing Alan M Turing, 'Computing Machinery and Intelligence' (1950) 59 Mind 433.

[127] Volokh (n 77), ibid.

Volokh's argument, read in light of these experiments, also offers an important response to concerns about judgement set out by Kerr and Mathen. Recall that for Kerr and Mathen, AI driven robots could not engage in important facets of judging sketched by Wittgenstein, Hart, and Dworkin — following a rule rather than merely comporting with it, assuming an internal point of view, and passing judgment as a member of a community with a holistic sense of its moral commitments. AI could not do so, in short, because it could not feel or experience a personal moral compulsion to decide a case one way or another, to care about whether it is framed within one moral framework or another. Volokh's analysis suggests, however, that in terms of capability, these concerns should pose no impediment to having AI decide cases so long as whatever AI does decide, its reasons are cogent and persuasive — to the point of being indistinguishable in quality from those of a human judge. So long as they are, it should not matter whether the entity making the decision arrived at it by merely comporting with rather than following a rule, assuming an external rather than an internal point of view, or passing judgement without having been a member of a community with a holistic sense of its moral commitments.[128] Put otherwise, if AI is capable of mimicking the output of a human judge, the only concern should be over whether a human would approve of or affirm that decision, not with whether they were the first to make it. The point then would be that while Wittgenstein, Hart, and Dworkin's concerns provide criteria for excellence or integrity in *assessing* a judgment, they should not be read as offering reasons for why only a human can *make* a good judgment.

The scholarship canvassed in Part I of this paper contained three further important concerns about AI replacing humans in legal judgement: AI's limitations in terms of discerning context and shifting social norms over time, and algorithmic bias and opacity. Recall that for Rebecca Crootof, legal judgement is a 'distinctively human skill' that requires being 'sensitive to context, both to extenuating circumstances in individual cases and shifts in social norms over time,' and an ability to 'flexibly apply legal rules.'[129] Ignacio Cofone had captured the consensus in a wide body of literature on algorithmic bias in making the point that AI's output in judgment — at that point in time, confined for the most part to predicting probabilities, outputting a score — would almost invariably reflect a bias of some kind in the algorithm or

---

[128] This is not, however, the same as saying that it would make no difference to the community if a cogent and persuasive decision were made by a machine rather than a human. Again, however, I set aside that question (should AI replace human judges) and focus instead here on whether AI would be capable of making a decision indistinguishable in quality from one made by a human.

[129] Crootof (n 39) 238.

training data.[130] He also aptly summarized widely shared concerns in the pre-2022 scholarship about algorithmic opacity in judgment. As noted in Part I of this paper, the weight of scholarly opinion had held that while programmers might take steps to reduce opacity, relying on AI to make legal decisions would likely always be fraught with suspicion about hidden assumptions and bias in the code. How well, if at all, can generative AI mitigate these concerns?

The experiment here suggests that using language models to make judgments can mitigate these concerns to a significant degree — but that, again, doing so depends on the way they are used. To be clear, the material on which a language model is trained may contain biases or reflect outdated social norms, and the manner in which it functions, how it decides normative questions, may remain opaque. Yet by asking a language model to decide a case by drawing on uploaded party materials and to provide extensive reasons for its decision, a model's output can reflect a sensitivity to context and shifting norms (as outlined or mooted in those materials), and it can mitigate bias and opacity by being transparent about its reasons for decision. By being asked to explain and justify explicitly, in plain language, why it is deciding one way or another, AI makes use of the primary method by which humans establish transparency, fairness, and objectivity in judgement: by offering a cogent and persuasive explanation of its reasons for a decision.[131]

Courts in Canada assume this logic in assessing whether decisions by an administrative tribunal are reasonable. In *Vavilov*, the Supreme Court held that 'a reasonable decision is one that is based on an internally coherent and rational chain of analysis and that is justified in relation to the facts and law that constrain the decision maker.'[132] The same court held in *Dunsmuir* that the 'reasonableness' of a decision 'is concerned mostly with the existence of justification, transparency and intelligibility within the decision-making process', and with 'whether the decision falls within a range of possible, acceptable outcomes which are defensible in respect of the facts and law'.[133] To assess reasonableness, the Court could point to no other criteria of merit *aside from* the cogency and persuasiveness of reasons for decision *on their face*.

---

[130] Cofone, 'AI and Judicial Decision-Making' (n 9) 8.

[131] Volokh (n 77) 1154.

[132] *Canada (Minister of Citizenship and Immigration) v Vavilov*, 2019 SCC 65 [85].

[133] *Dunsmuir v New Brunswick*, 2008 SCC 9 [47].

But this is not to say that fairness or transparency in judgment — which is to say, legitimacy in judgment — turns exclusively on what a decision or opinion contains on its face. A well-reasoned decision might still be considered unfair or not fully transparent if it turns out that the decision-maker sided with one party without disclosing a hidden social connection that later comes to light. Judgments in law must not only be made impartially, but *appear* to be so.[134] As suggested earlier, if it turns out that a language model makes a well-reasoned decision that reflects a deeper technical bias (that may or may not be gamed by a party), being cogent and persuasive may not be enough to meet the need to appear fair and impartial.

However, this concern should not be overstated. Short of a clear ability to game or hack a model's output in one's favour, the fact that AI's processes remain opaque or carry some measure of bias in their operation should not in itself pose a hindrance to making effective judgments — since the same is true of humans. In assessing the merits of decisions by trial courts or tribunals, appeal courts are not concerned with the judge's mental health, internal point of view, or personal investment in the community within which he or she is judging. They are not concerned with *how* — physically, mentally, morally — a judge arrived at a decision but rather with the reasons offered for *why* they did so. Analogous reasoning applies to AI. Internal processes that remain mysterious and opaque may play a role in producing a decision. But short of evidence that a model's decision-making process can be readily gamed or is fraught with significant bias, what should matter in assessing AI's capability to make legal judgements is the quality of the reasons it provides to support them.

Reservations and conclusions

The public release of ChatGPT in 2022 pointed back to an unexpected innovation in 2017. Researchers at Google came up with an approach natural-language processing involving machine-learning that brought about a quantum leap in AI's abilities to read and write.[135] It soon enabled AI to process text, argument, and expression in a manner and to a level of

---

[134] A canonical formulation of this principle can be found in Lord Hewart CJ's dicta that: '[it] is of fundamental importance that justice should not only be done, but should manifestly and undoubtedly be seen to be done': *The King v. Sussex Justices, Ex parte McCarthy*, [1924] 1 KB 256, at 25. In the Supreme Court of Canada's decision in *R v S (RD)*, [1997] 3 SCR 484, Cory J held at para 91: 'A system of justice, if it is to have the respect and confidence of its society, must ensure that trials are fair and that they appear to be fair to the informed and reasonable observer. This is a fundamental goal of the justice system in any free and democratic society.'

[135] Ashish Vaswani and others, 'Attention Is All You Need' (2017) 30 *Advances in Neural Information Processing Systems*; Steven Levy, '8 Google Employees Invented Modern AI. Here's the Inside Story' (*Wired* March 2024) <https://www.wired.com/story/eight-google-employees-invented-modern-ai-transformers-paper/> accessed 7 April 2025.

sophistication not envisioned in the scholarship on AI in judgement as late as the end of the 2010s. AI's role in judgement can now vastly exceed the predictive, score-generating function that formed the focus of the earlier scholarship. Unikowsky's findings attest to this. The thrust of his experiment — showing a language model could decide 27 of 37 recent US Supreme Court cases the same way the Court did and offer persuasive reasons in the remaining 10 — shifts the ground for assessing AI's role in legal decision-making. If we assume that the ability to decide cases should be assessed in terms of the ability to provide cogent and persuasive reasons for decision, generative AI can be used in a way that mitigates many of the concerns about AI in the earlier scholarship.

But certain limitations remain, and they become especially salient at the trial level. Recall that for Tania Sourdin and Richard Cornes, judging is 'multifaceted.'[136] It can involve activism, 'complex interactions with people,' 'social commentary,' 'emotional understanding of the parties,' and more.[137] They also argued that the 'social legitimacy' of a decision rests in part on the perception of it being the outcome of 'a judicial decision maker struggling conscientiously, in public, with the detail of a case'.[138]

The findings of the Unikowsky experiment involving generative AI in appellate judgment might be extrapolated here to the trial context, but Sourdin and Cornes' concerns point to AI's capacity at the trial level being more limited at present. A language model might play a useful role in assisting a trial judge in producing a draft decision based on party materials, a judge's own summary of the evidence, and a judge's prompting as to how the case should be decided. But trials mostly involve listening to, watching, and managing people — in distinction to appeals, which mostly involve reading, analyzing, and responding to written texts. A literature is now emerging that explores the potential role of new AI tools involving sound and images to assist judges in making findings of fact and to help mediators or arbitrators in various ways.[139] But these issues are beyond the scope of this paper.

---

[136] Sourdin and Cornes (n 21) at 3.

[137] Ibid 3 and 12.

[138] Ibid 12 (citing Micheal Kirby, 'The future of courts—do they have one?' (1999) 8 *J Judicial Adm* 383) and 13.

[139] Hannes Westermann, Jaromir Savelka and Karim Benyekhlef, 'LLMediator: GPT-4 Assisted Online Dispute Resolution' (arXiv, 31 July 2023) <https://arxiv.org/abs/2307.16732 accessed 8 April 2025>;Yiyang Mei and Michael J Broyde, 'Don't Kill the Baby: The Case for AI in Arbitration' (August 2024) SSRN <http://dx.doi.org/10.2139/ssrn.4936638>; David L Evans and others 'Dispute Resolution Enhanced: How Arbitrators and Mediators Can Harness Generative AI' (2024) 78(1) *Dispute Resolution Journal* 57; and Katie Shonk, 'AI Mediation: Using AI to Help Mediate Disputes' (Harvard Law Program on Negotiation March 2025)

At this point, we have enough evidence about AI's capabilities with language and legal reasoning to infer that language models can assist judges with crafting an outline or draft of a decision at all levels of court. Models have continued to improve and increase in ability and effectiveness. How much further they will progress is unclear, but the general direction of movement would seem to be inevitable. AI is proving to be more useful and efficient at more facets of the process of judgment, from research to consolidation (of evidence and precedent) to opinion drafting. This points to a future in which the process of legal decision-making will involve more automation and delegation; more of a shift from gathering and processing material to carefully prompting AI to act on that material and editing its output.

The evidence of AI's expanding capability canvased here also invites speculation in a broader sense. Just as computers and networks have made research, writing, and communication more fluid, automated, and immediate based on what they make possible, so too will AI likely transform judgment to be faster, more the product of automation, and more a combination of human intuition and computation. It points, at the least, to a situation where the line between the two becomes increasingly harder to discern — making the question of whether AI *should* replace human judges increasingly moot. As AI becomes more capable and plays a greater role in judgment, taking over more facets of it, the question may well shift from whether AI should replace humans in some broader sense to whether AI's judgment in any given case is acceptable and what it would take to make it so.

<https://www.pon.harvard.edu/daily/mediation/ai-mediation-using-ai-to-help-mediate-disputes/> accessed 7 April 2025.