Jus Cogens
https://doi.org/10.1007/s42439-026-00120-6

ORIGINAL ARTICLE

Objections to the Language Model as Judge: Normativity, Aristotelian Phronesis, and Arendtian Common Sense

Robert Diab[1]

Received: 4 December 2025 / Accepted: 20 April 2026

Abstract

Scholarship and policy on the use of artificial intelligence in legal judgment continues to be shaped by a consensus formed at an earlier stage in AI's development. AI cannot be sensitive to context, reason normatively, or be responsive to novel facts. Recent experiments challenge these assumptions by showing that language models prompted to decide a case based on party materials uploaded to the model can render a reasoned decision comparable in quality to that of a human. This new capability of AI casts in a new light philosophical reservations about algorithmic judgment based on ideas about normativity, Aristotle's concept of *phronesis* (practical wisdom), and Hannah Arendt's theories common sense and reflective thought. Revisiting these arguments in the context of AI's new capabilities illuminates how differences between human and algorithmic judgment are evolving and supports a pragmatic alternative basis for accepting the language model as judge in some cases.

Keywords AI · Language model · Judgement · Automated decision-making · Arendt · Aristotle

1 Introduction

Theoretical and legal scholarship on the role of artificial intelligence (AI) in judgment is extensive and reaches back many decades (Surden 2019; Cofone 2021). A common feature across the literature well into the present is a focus on forms of AI or algorithmic judgment designed to output scores or probabilities: a risk of recidivism, credit worthiness, the likelihood of success in litigation (Gouge 2021; Raso 2021; Beatson 2018). Focused on these earlier forms of AI, a consensus has taken shape in

✉ Robert Diab
rdiab@tru.ca

[1] Faculty of Law, Thompson Rivers University, Kamloops, Canada

Published online: 07 May 2026

scholarship and policy that AI lacks the capacity to carry out essential functions of legal judgment. AI cannot be sensitive to context, reason normatively, or be responsive to novel facts. It might assist humans in the task of judgment, but in high-impact cases involving liberty or property interests, only human judgement is effective and legitimate.

Recent experiments challenge these assumptions by showing that language models prompted to decide a case based on party materials uploaded to the model can render a reasoned decision comparable in quality to that of a human (Unikowsky 2024a, b ; Gandall et al. 2025; Posner and Saran 2025; Diab 2026). Models drawing on party materials demonstrate an ability to be attentive to law and context; to reason normatively; to craft new legal rules; and to apply them to new facts. AI involving a language model is capable, in short, of passing a version of a judicial Turing test, as anticipated in earlier scholarship (Volokh 2019; Susskind 2019), consistently approximating the actual output of apex courts when tested with factums from those cases (Unikowsky 2024a, b) or outcomes in arbitration (Gandall et al. 2025).

This evolving capacity of AI is only beginning to be explored and has yet to register among scholars and policy-makers. This paper highlights one dimension of the results of these recent experiments: the way in which a language model outputting a reasoned decision in response to party materials overturns earlier assumptions that AI was limited to applying a simple form of if-then reasoning or was otherwise confined in its decision-making to employing discrete algorithms that could not be applied to new and unpredictable scenarios. Instead, language models bring about a new form of algorithmic judgment that mimics aspects of a human's ability to think through a problem in natural language.

This new form of artificial reasoning in judgment gives cause to revisit a series of notable philosophical objections to AI's role in judgment. The paper highlights three. One is that regardless how effective, correct, or persuasive AI decisions might be, they could never be a legitimate substitute for human judgment in a liberal political order because AI decides cases on a statistical rather than normative basis, thus reflecting an irreducible arbitrariness (Afrouzi 2024; Tasioulas 2023; Grimmelmann et al. 2026). A second argument holds that because AI decides cases algorithmically, it cannot possess *phronesis* or practical wisdom, as Aristotle conceived this (Groff and Symons 2024). A third argument claims that AI decisions cannot reflect an essential element of judgment in a democratic community, inter-subjective common sense and reflective thought as Hannah Arendt conceived it (Herzog 2021; Tajalli 2021).

The paper reevaluates these arguments in the context of AI's evolving capacity for judgment, noting various ways this puts pressure on the distinction between human and algorithmic judgment. The aim is not to persuade the reader that new capabilities refute the philosophical objections canvassed here, but rather that these capabilities cast the objections in a new light. They remain true but with important qualifications that suggest a shortening of the distance between a human and an AI judge. With these qualifications in mind, the paper closes by advocating the use of language models as aids to or substitutes for human judgment in certain cases, such as those involving consent of the parties. The argument for AI's utility would rest, in those cases, on a pragmatic basis: that AI decisions should be assessed in terms of their cogency and persuasiveness, rather than the process underlying their production.

1.1 Earlier Perceptions of AI's Capacity for Judgment

The scholarship on AI in legal judgment was well developed prior to the advent of ChatGPT in 2022 (see the survey in Cofone 2021; Surden 2019). The common thread is its focus on forms of AI that carry out the limited function of predicting a score or probability based on pattern matching or similarities among defined variables. These include tools for predicting recidivism, deciding immigration eligibility, detecting tax evasion, or allocating social benefits (e.g., Gouge 2021; Raso 2021; Beatson 2018). The limited scope of these tools gave rise to three common assumptions about why AI lacked the capability to replace humans in the task of legal judgment.

One was that AI involving forms of machine learning of any complexity concealed the technical process by which a tool arrives at a decision, making it impossible to verify the process was not biased or gamed or otherwise unreliable. As Amy Gouge writes: "[w]ithout being able to identify the factors that the AI considered through a process of autonomous self-learning, it would be impossible to communicate the reasons for the AI's outcome or to ensure the AI's reasoning process accounted for all relevant factors" (2021, p. 29; see also Surden 2019; Conglianese and Lehr 2017). A closely related concern was that AI tools could not provide reasons for a decision that would "engage with the specifics of an individual case," thus failing to provide a nuanced or factually-responsive justification for a given outcome (Daly 2023, p. 10).

More broadly, AI lacked the capacity to engage in forms of reason thought to be essential to legal judgment. AI lacked "common sense, empathy, or moral reasoning" and could not effectively apply "discretionary rules that demand appreciation of context" (Gouge 2021, p. 29, citing Surden 2019, p. 1309). As Coglianese and Lehr argued:

> Machine-learning algorithms cannot directly make the choices about [] different aspects of a rule's content not only because some of these choices are normative ones, but also because learning algorithms are merely predictive and thus unable to overlay causal interpretations on the relationship between possible regulations and estimated effects. (2017, p. 1173; see also Sunstein 2001.)

AI could not "apply legal rules in accordance with changing social mores" (Crootof 2019, p. 237); "machine learning techniques" were thought to be "only useful where analysed information is similar to new information presented to the AI" (Sourdin and Cornes 2018, p. 99). Many agreed that AI "will founder when applying an ambiguous rule to a novel or complex situation" (Crootof 2019, p. 239; Shay et al. 2016, p. 274; Coglianese and Lehr 2017, p. 1117). Writing as late as 2019, Surden asserted: "AI tends to work poorly, or not at all, in areas that are conceptual, abstract, value-laden, open-ended, policy or judgment-oriented" (pp. 1322-3).

1.2 AI's Evolving Capacity for Judgment

The concern in earlier scholarship that AI tools based on machine learning are irreducibly opaque remains current. There is some debate as to the extent to which outcomes grounded in machine learning are explainable if not fully interpretable (Zhao

et al. 2024; Maruthi et al. 2022; Arrieta et al. 2019). Yet AI involving large language models has brought about a host of new capabilities that call other assumptions canvassed here into question: that AI cannot give reasons that justify or render the basis for arriving at a decision intelligible; that without this capability, AI decisions are unreliable due to the possibility of bias in the underlying code; and that AI cannot reason judiciously (be contextually sensitive, engage in normative reasoning, or apply law to individual facts correctly). Scholars and jurists have begun to explore AI's evolving capacity by conducting experiments with language models in the appellate, administrative law, and arbitration contexts (Unikowsky 2024a, b; Gandall et al. 2025; Posner and Saran 2025; Diab 2026; and Daly 2023). The experiments demonstrate that a language model prompted to act upon party materials uploaded to the model (factums, briefs, written argument) is capable of producing an opinion that sets out cogent and persuasive reasons that justify a decision. Models can make normative judgments and engage in other forms of creative legal thought, such as formulating new legal tests. In a critical assessment of these experiments, Grimmelmann et al. (2026) concede that "LLM output now replicates rational, eloquent argumentation that applies precedent to novel facts. An LLM can produce text that may be formally indistinguishable from—or even formally superior to—the reasoning described by an opinion written by a human judge" (p. 286).

The reader is invited to consult the sources noted above canvassing these experiments in detail. For the purposes of this paper, I briefly describe the method and findings of three of them. In the summer of 2024, Adam Unikowsky uploaded to Anthropic's Claude 3.0 language model briefs from 37 of the United States Supreme Court cases in the current term and asked it to briefly outline a decision in each case in three to four paragraphs. He found that Claude decided 27 of the 37 cases the same way the Court did and that in the remaining 10, he "frequently was more persuaded by Claude's analysis than the Supreme Court's" (2024b). His write-up of the experiment includes a detailed discussion of six decisions rendered in the previous week (Unikowski 2024b). Claude, he writes, "nailed five out of six [of the cases], missing only *Campos-Chaves*, in which it took the dissenters' side of a 5−4 opinion, which is hardly 'wrong.'" Unikowsky documents Claude's ability to formulate different and more elaborate legal tests in relation to those found in the actual decisions—and to apply these tests effectively to the facts in the case at bar. The experiment as a whole presents strong evidence of a language model's capacity to consistently render the outline of a decision in an apex court case that approaches the quality and sophistication of the court's actual ruling.

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

itself (see the detailed comparisons in Diab 2026). The reasons in each case also demonstrated an ability to craft a new legal test or standard and to apply it to novel facts.

Gandall et al. (2025) conducted an experiment with software that two of them had created called Arbitrus.ai to demonstrate that it "performs just as well as human arbitrators right now" (p. 1) by giving it 100 hypothetical scenarios involving party materials created by another language model. The authors found that "Arbitrus.ai performed as expected: it did not hallucinate; it answered the issues in controversy, and almost always—with two narrow exceptions—grounded those answers in relevant case law" (p. 56). The two exceptions involved the model drawing on cases with "disanalogous facts," amounting to an erroneous reliance on certain earlier cases; errors that a human judge might readily have made.

A common finding across these experiments is that a language model can be responsive to party positions, be contextually sensitive, and can apply law correctly to novel facts if provided materials containing these specifics. The most significant advance of generative AI, however, is its ability to demonstrate sophisticated legal reasoning, or at least the appearance of this, to *justify* a decision. What the models do not do is reveal the specifics of the technical process, under the hood, for reaching their output. Models provide a rational basis for a decision, but not its technical or actual basis in a given case. This distinction lies at the heart of the philosophical concerns with language model judgment to be explored in the next segment of this paper. The point to highlight here is that if provided sufficient material, a model *can* generate reasons for decision that offer a transparent and intelligible explanation as to why a certain outcome should be accepted as correct in a legal and moral sense—in the absence of evidence showing that underlying processes were biased or somehow flawed.

Before turning to these philosophical objections, I note that debate continues to unfold over how consistent and reliable language models are when asked to play the role of judge. Jonathan Choi (2025) has demonstrated that "LLM judgments are highly sensitive to prompt phrasing, output processing methods, and model training choices," leading to different outcomes and varying degrees of confidence in them based on slight changes in the wording of a prompt (p. 1). Yet Kieffaber et al. (2025) have argued in response that Choi's concerns apply to "off the shelf" language models, but not to bespoke models that incorporate forms of machine learning known as "classifiers." In the legal context, classifiers can help models make more consistent judgments about factual scenarios. They do so by making decisions about them not based on next-token predictions alone but also by identifying categories into which a factual scenario belongs. A classifier helps a model decide cases by assessing what *kind* of case the facts at issue are more like than any other, and how courts have tended to treat that kind of case. The authors demonstrate that language models that incorporate classifiers generate output in relation to new factual scenarios with a high degree of consistency and accuracy in terms of the law despite a wide variation in prompt formulations.

The upshot of Kieffaber et al's technical innovation involving classifiers is that while popular models such as Claude or ChatGPT may be unreliable for legal judgment, models can be modified to allay concerns about arbitrariness in judgment on grounds of inconsistency. However, other philosophical concerns remain.

2 Philosophical Objections to the Language Model as Judge

The first part of this paper canvassed assumptions about AI's functional limitations that rendered it incapable of replacing humans in legal judgment. Recent experiments show that many of these assumptions about AI are no longer true; its functional capabilities have evolved. The question in what follows is whether AI's new abilities in judgment, involving certain ways of using language models, can overcome philosophical reservations about automated judgment. If AI *can* perform as well as a human judge in some contexts—such as appellate judgement or mediation—are there good reasons to conclude that it *should* not replace a human?

Among the many theoretical arguments made against full delegation of legal judgment to AI (see the surveys in Surden 2019; Coglianese and Lehr 2017), this section revisits three bodies of critical commentary: those concerned with normative versus computational reasoning, Aristotle's notion of *phronesis*, and Hannah Arendt's ideas about the role of common sense and reflective thought in judgment. The aim here is to consider how AI's new capabilities provide a basis for complicating or at least qualifying these arguments — but not for refuting them. In ways to be seen, important differences remain between human and algorithmic judgment. New AI capabilities may help to illuminate the evolving nature of these differences.

i. Normativity

One common thread in the literature argues that while AI might someday be capable of engaging in normative reasoning, it cannot decide a case on a normative basis—it cannot choose a given outcome because law or morality compels it. A computer can comply with a rule but not follow it. Only a human can do this. And since our sociocultural assumptions about what makes a legal decision fair and just rest on the idea of doing what law or morality compel in a given situation, only a human can play the role of judge legitimately in high-stakes cases (i.e., those involving liberty, equality, property).

Although a host of authors have advanced a variation of this argument (Hildebrandt 2008; Berman 2018; Williams 2022), I single out two notable recent iterations. Ebrahimi Afrouzi (2024) has offered what might be the clearest and most extensive defence of it. His point of departure is to assume that even if a model's underlying processes were fully interpretable and free of bias, the model could still not render a decision we could accept as a legitimate substitute for human judgment because its basis would be found in nothing more than "statistical correlations" (p. 370). Regardless of the reasons it might offer to explain a decision, an AI judgment would be "necessarily deficient in rationale" because its true basis would lie in technical processes that are algorithmic or mathematical rather than normative (ibid.). A rationale is a reason offered in response to a normative inquiry, an explanation that might be as simple as 'X should be done because a rule, precedent, or other legal value compels it.' As Afrouzi writes, "legally valid decisions must be correct not only in their holding but also in their rationale, and yet, AI decisions cannot even hope to be correct in their rationale." AI could thus never provide the "right kind" of justification required in law (p. 384).

Similarly, John Tasioulas (2023) has argued that a decision based on machine learning involves a "categorically different process from the essentially normative enterprise of justifying a decision in a particular case by reference to the relevant reasons for that decision" (p. 14). An automated decision is one that might comply with a rule but does not follow it; and as a consequence, AI cannot provide assurance that it was not just "coincidentally congruent with the law", that it was "arrived at precisely because it is in accord with the law" (ibid, p. 13). The reasons that a language model might provide for a decision in its textual output are merely *ex post* justifications for it. While they might support the decision, they cannot truly justify it since they were not "causally efficacious" in bringing it about (ibid, p. 15).

One rejoinder to these arguments is that humans might reason in the same way, deciding upon an outcome first and then formulating reasons to justify it. Affrouzi's response is that at least in theory, we can reason to a result by first asking what law or morality require and choose an outcome on that basis—providing a causal connection (ibid., p. 387). When a human offers an outline of their reasoning process, it may justify the decision because it is the real basis for the outcome, or must be accepted as such in the absence of evidence to the contrary. With language models, this can never be the case. Yet some authors suggest the distinction here is illusory. At a neurophysiological level, humans also make decisions for reasons that are either opaque or non-normative. As Eugene Volokh asserts: "[i]f we are honest with ourselves, we often can't really tell with confidence why we reached a particular judgment […] We have reactions because of the real neural nets in our brains, and then we can offer explanations that we hope persuade" (2019, p. 1165; see also Berman 2018, p. 1319, noting "contexts where human decision-making is itself opaque"). Affrouzi's response is to argue that human decisions are not reducible to processes on which consciousness "supervenes," but are instead made "at the level of conscious human reasoning"—a level to which AI has no equivalent (2024, p. 387).

The distinction being drawn here between a statistical and normative basis for judgment has provided a grounding for various broader arguments about automated judgment being incompatible with the rule of law or with democratic ideals (e.g., Berman 2018; Re and Solow-Neiderman 2019; Frost 2024). Some authors contend that AI decisions are incompatible with values or concepts of what makes for legitimate legal judgment in a self-governing social or political order (Affrouzi 2024; Daly 2023; Tasioulas 2023). Others contend that AI decisions are unlikely ever to be accepted as legitimate due to widely held socio-cultural perceptions about machines as vacuous or merciless in their determinations (Kahneman et al. 2021; Re and Solow-Neiderman 2019).

A recent paper critical of AI's capacity for judgment combines Affrouzi and Tasioulas's concerns with normativity and broader arguments about democratic values—and does so in response to the experiments canvassed above by Unikowski, Keiffaber et al., and others. Grimmelmann et al. (2026) argue that to "conflate persuasiveness and coherence [in judgement] with authority and accuracy is a category error" (p. 286). To be perceived as legitimate, adjudication in law must conform to implicit "procedural criteria" that include a decision-maker acting in good-faith, impartially, and rationally (ibid., p. 287). A language model's output might resemble legal reasoning and it may arrive at a result that is correct in law, but it would not have done so

as a result of "inductive or deductive approaches" (ibid, p. 289). The assumption that human judges engage in this form of reasoning reflects the notion that "legality is a social fact"—a function of being associated with a "social process of adjudication" (ibid., p. 289, 292). The authors illustrate this by pointing to the role of juries, the value of which lies not in the binary output they produce but in the "deliberative process that legitimates [the jury's] authority" (ibid, p. 291). Similarly, a judge's authority and the legitimacy of the reasons they provide in support of a decision derives not just from the reasons being coherent and persuasive but from the perception that the judge acted "fairly, objectively, on the basis of the evidence, and in accordance with the relevant authorities" (ibid., p. 308). This expectation is something a machine can never satisfy, since its processes are merely algorithmic or computational.

Grimmelmann et al. (2026) repackage two arguments against the language model as judge noted earlier. AI can only decide for correlative rather than causal or normative reasons; AI can only conform to a rule rather than follow it. And only an entity which can do the latter can be accepted as a legitimate source of legal judgment. Eugene Volokh has offered the strongest counter-argument against both claims: i.e., that AI can only be coincidentally correct in its legal judgments (given its basis in statistical, correlative processes rather than normative, causal reasoning) and therefore that persuasiveness is not enough to render AI decisions legitimate. His argument is that the only way of knowing whether a given outcome is normatively correct—is what must follow from earlier legal doctrine or ethical principles applied to the evidence in a given case—is persuasion itself (Volokh 2019, p. 1161). This is really a way of saying that since correctness in law or morality is a matter of persuasion all the way down, people ought to overcome their antipathy and accept the idea of the AI judge as legitimate because AI is just as effective at doing the only thing that humans really do: produce cogent decisions that persuade us that they are normatively correct.

This may be a powerful argument in favour of using AI to assist in deciding certain cases or to decide them alone. A human judge, for example, might choose to endorse the most persuasive decision among two or three opinions that a language model has produced, and do so despite the underlying statistical or algorithmic rather than normative basis for the opinion. Or parties that have consented to their dispute being resolved by AI alone might be satisfied with the reasons for decision that it offers due to their cogency and persuasiveness. But this is not an argument in response to the question of legitimacy. Would this form of judgement be perceived as legitimate if *imposed* on litigants?

Grimmelmann, Sobel, and Stein are doubtful about this. Even if one concedes that AI can play the role of judge as well as a human, they argue, it is unlikely that machines will ever gain acceptance for high-impact decision-making given our cultural attachment to the social dimension of judgment—epitomized by the role of the jury, but also present in humanistic conceptions we harbour about the ideal judge as a *person* who is empathetic, reflective, and impartial. A judge guided by AI's suggested outcomes would lack impartiality; a decision made by AI alone would not escape the appearance of arbitrariness. These authors may well be correct; we may never overcome our antipathy to AI judges for lacking these qualities. But both Volokh and Suskind caution against being too quick to assume this. As Susskind (2019) notes, "it is probable that our grandchildren will have different views from ours" on "the

computer judge" since "they will live in an age when it will be commonplace for machines to be unarguably superior to humans in many walks of life" (p. 292; Volokh 2019, p. 1171). This points to at least the possibility of AI gaining acceptance for a role in judgement in some contexts.

ii. Phronesis

Scholars have been critical of AI's possible role in legal judgment from another angle: its inability to possess *phronesis* or practical wisdom as Aristotle conceived it in the *Nicomachean Ethics*. Although a body of scholarship has emerged bringing Aristotle's ethical thought to bear on AI (e.g., Eisikovits and Feldman 2022; Sullens 2021; Tsai and Ku 2025), I single out the arguments of Groff and Symons (2024) for the depth of their engagement with *phronesis* specifically in relation to language models.

Briefly, for Aristotle, *phronesis* involves a cognitive ability to decide what to do in unique, unprecedented situations, those to which no prior rule or principle readily applies (1941, 1040a, 1142a). *Phronesis* is a component of "virtue in the strict sense" (1144b), along with *hexis*, or good character, which involves responding to situations with the right emotions and having a desire to do good when faced with certain choices (ibid., 1144a; 1105b). A person with *hexis* might demonstrate a kind of "moral virtue" by making choices in difficult situations that happen to be virtuous, but they may not draw on practical wisdom in making these choices. Only a *phronemos*, a person possessing practical wisdom, has the capacity to know when a rule should be broken or what justice calls for in a specific or novel situation, by drawing on years of practical experience (ibid., 1143a).

Groff and Symons (2024) contend that while a language model might possess moral virtue—in the sense of making morally acceptable choices—it could never possess *phronesis* or the practical wisdom to decide what to do in new situations that engage conflicting moral principles (p. 225). They infer that language models lack this ability by asserting that AI based on machine learning would have only two ways of demonstrating moral virtue or making moral judgments. One is to train an AI on a large set of examples of moral problems and responses to them considered virtuous (ibid., p. 224). The other approach is to "add moral (or legal) heuristics as training data after the LLM is in place as a way to censor the outputs of the model in accordance with commonsense heuristics" (ibid., p. 225). These would dictate that the model should not lie or pretend to have a body or should avoid stereotypes. The censorship would operate by testing potential outputs against a set of examples of each rule in the training data. Whether something amounts to a stereotype, for example, would depend on its similarity to what others have judged to be a stereotype.

The problem that a language model cannot surmount, Groff and Symons argue—what hinders a model from possessing *phronesis*—is that "AI will only ever be capable of actions derived from a probability distribution over elements of [a given] set or via heuristics or rules that serve as the basis for training" (ibid, p. 225). "The fundamental problem," they assert, "is that, if Aristotle is right, there is no rule, or set of conditional if-then sequences that could be coded, for telling what the wise thing to do would be in every possible situation" (ibid). AI that operates on pattern-matching cannot decide what to do in situations involving "multiple or even competing moral

principles, requiring the ability to weight different factors and trade-offs" (ibid., p. 226). For example, how would AI decide in a difficult case whether honesty should prevail over kindness? There is, they contend, "no such thing as programmable practical wisdom as per Aristotle"; no way that a computer could be "capable of judging between competing goals in a nonarbitrary way" (ibid, p 227). Practical wisdom requires an ability to decide which values must take precedence over others in a given situation. How, they ask, could an AI be trained to do this? How could it carry out "multiple kinds of optimization tasks without there being some master optimization task that would automatically subordinate the other two"? (ibid, p. 228).

When language models are presented with a moral dilemma (in the form of a legal dispute), do they decide what to do by relying on a form of pattern-recognition or 'if-then' reasoning? One inference we might draw from the experiments canvassed above is that language models drawing on party materials to resolve a dispute by engaging in a form of if-then thinking but not that alone. Models come up with an answer to a moral or legal quandary by predicting which tokens in a sequence should follow the initial sketch of the problem (in the prompt and in opposing party materials). In other words, a model draws on probabilistic correlations derived from text in the training corpus coupled with the facts in the problem in question, along with arguments for opposing outcomes. A model thus decides which principle should prevail (be honest or be kind) not on the basis of applying a master principle but by predicting what humans would *say* about a given problem based on what they have said over a large corpus of material involving similar words or ideas. On this view, generative AI does not possess *phronesis* but it comes closer to demonstrating it than Groff and Symons' argument would allow.

But skeptics of artificial *phronesis* might point to Jonathan Choi's findings (2025) about prompt sensitivity shaping differing outcomes to raise a doubt about whether models are really deciding cases at all—rather than predicting a sequence of tokens based on how a prompt is worded, which party's brief was uploaded to the model last, which one was longer or better, and so on. On this view, language models drawing on party materials demonstrate nothing close to moral virtue, let alone true practical wisdom. They simply generate the outcome favoured by the wording of a prompt and the better factum or brief on which the model draws. And Kieffaber et al's solution (2025) involving 'classifiers' is of no assistance here, since that involves a fall-back to pattern recognition, which runs into the basic problem that *phronesis* is meant to solve: what to do when there is no pattern into which a fact scenario falls.

Two arguments in defence of artificial *phronesis* might be offered to these powerful objections. The first is that even where models do not rely on classifiers, prompt sensitivity (along with the quality and order of party materials uploaded to the model) may lead to a wide variation in outcomes and a degree of inconsistency in judgment—i.e., when a model is asked several times to decide the same case using a different set of prompts or materials. But models are still capable of deciding novel cases and producing cogent and persuasive reasons without relying on a master principle; that is to say, they can resolve novel moral or legal disputes relying primarily on next-token prediction. And as the experiments canvassed above demonstrate, the output can closely approximate in quality and result the actual decisions of human judges. (As noted earlier, Unikosky had found Claude 3.0 decided 27 of 37 cases the same

way the US Supreme Court did that year, and decided the remaining 10 cases with reasons he found more persuasive than the output of the court itself.) This does not overcome concerns about prompt sensitivity and thus a certain kind of arbitrariness in judgment, but it does suggest that AI can demonstrate something like *phronesis* without relying on a master principle to settle disputes involving conflicting principles.

But what about where a language model does rely on classifiers to bolster consistency? Does this not prove that models are confined to the level of applying principles to similar cases? Does it not prove that without a master principle to resolve cases involving a conflict of principles, AI cannot possess practical wisdom? One might argue that it proves the contrary. It readily invites the inference that *phronesis* or practical wisdom is nothing more than the ability to discern in a given situation which kind of case these novel facts are *more like* than any other. The *phronemos* decides that in this situation, it is better to be honest than to be kind by discerning that the facts in this case are most like the ones in past cases where it worked out better to be honest than to be kind and vice versa. On this view, *phronesis* does not require possession of a master principle that would settle the conflict of principles in a given scenario, but rather an ability to discern which of the two principles should prevail on the basis of the most plausible identification of the scenario as *essentially* one kind of case or another (e.g., an instance of the 'it is better to be honest' principle).

This would not, however, settle the question of whether a language model equipped with a classifier would possess *phronesis* or demonstrate it effectively, because an AI classifier would make determinations based not on human intuition but on statistical probability. It would decide that the scenario in question is more like cases in which it was better to be honest than kind based on a correlation of the variables it happens to measure. The judgment, in other words, may be flawed compared to that of a truly wise human. But it would be difficult to test this proposition, since the ultimate determination of whether a scenario is *essentially* more like one category of cases than another is not susceptible to measurement or verification. There is no basis on which to test whether a human makes this determination more accurately or effectively than a machine. We fault the machine for doing it on the basis of statistical or mathematical measurements, but we might be doing something similar subconsciously.

iii. Common Sense and Reflective Thought

A third strain in the philosophical criticism of the possible role of AI in judgment relates to common sense and reflection. Scholars have drawn on Hannah Arendt's writing on moral judgment—her arguments about its intersubjective quality and the importance of thinking to the process—to cast doubt on the prospects for using AI as a viable substitute for human decision-making (Herzog 2021; Tajalli 2021). Judgment, for Arendt, involves an appeal to common sense because moral decisions are not made by following rules but by imagining what other members of a given community would consent to or support being done in a specific case. AI based on opaque algorithms or machine learning, some argue, cannot engage in this imaginative and reflective process, and thus it can neither persuade us nor assume responsibility for its judgment in a way that a human judge can.

Before turning to the question of how AI involving language models complicates this strain of criticism, it may help to expand briefly on Arendt's ideas about judgment—to be more precise about the ability that AI is assumed to not possess. Arendt's theory of judgment was unfinished at the time of her death, and scholars debate whether and how her arguments in various essays cohere into a single theory (Tömmel and Passerin d'Entreves 2025). I sidestep this debate by drawing upon Bryan Garsten's (2007) careful reconstruction of Arendt's ideas about intersubjectivity and common sense in making moral judgments for insight into thought processes underlying legal judgment in difficult cases.

An early point of departure for Arendt in her thinking about judgment was her inclination to see the reliance of Nazi official Adolf Eichmann on rule-following as revealing a flaw in Kant's moral theory (Garsten 2007, p. 1077). Eichmann argued that he should incur no liability for his role in the Holocaust because he acted under a duty to follow the law; he subordinated his will to prevailing legal and moral imperatives. In Arendt's view, this was consistent with Kantian morality, which sought to make ethical conduct a matter of complying with an obligation or imperative, of subordinating the will to reason, on the assumption that obedience was a virtue (Arendt 2003b 1966, p. 72). But in a secular world that valourizes individual dignity and responsibility, obedience cannot be a virtue (Arendt 2003a 1964: 48). Making moral judgements, for Arendt, entails an ability to "arbitrate between reasons without being subject to them" (Arendt 2003b1966, p. 131). How we do this involves a degree of "spontaneity" and "mystery" (ibid.). But this raises the question of how a moral decision that does not follow an imperative would not be purely subjective and only moral in the Nietzschean sense of expressing a free will not constrained by societal mores (Nietzsche 1990; cited in Garston 20074, p. 1084)

Arendt's solution to this conundrum—moral judgments being "neither objective and universal nor subjective, depending on personal whim"—was to conceive of them as "intersubjective or representative" (Arendt 2003b, p. 141). We make moral judgments not by following higher order rules or by deciding arbitrarily, but by imagining how other members of our community might perceive an issue and by choosing an outcome that we hope they would accept as valid (ibid., p. 144). We are guided in this regard not by resort to a general or abstract rule, but by reference to an exemplar or particular instance that "becomes valid for other particular instances"; her examples being Solon for insight and Bonaparte for leadership (ibid.). An exemplar differs from a general rule or schema by providing a *qualitative* basis on which to draw a comparison: a basis to ask not simply whether one is an instance of the other, but how well one approximates the *excellence* of the other. The legitimacy of our judgments, the inclination of others to agree with our qualitative assessments, will, however, depend ultimately upon "our choice of company" (ibid, p. 146−47). There is an inescapable degree of subjectivity and arbitrariness in this choice: why is *this* the group to be persuaded? Yet the decision that gains acceptance is neither a mere reflection of the group's norms and values, or a need to comply with them, nor one that takes "only myself into account" (ibid., p. 141). It lies between the two, seeking legitimacy through an appeal to consent, while being something for which we are individually responsible.

Arendt's primary concern with moral judgment was not that some might choose the wrong exemplars or the wrong company, but rather that some would demonstrate

an "unwillingness or inability" to do either and thus to engage in a "very common modern phenomenon, the widespread tendency to refuse to judge at all" (2003b, p. 146). This points in the direction of some of AI's functional limitations for legal judgment. For Lisa Herzog, decisions made by opaque algorithms, or forms of AI that do not reveal their underlying processes, circumvent the "mutual adjustment of perspectives" involved in Arendt's intersubjective conception of judgment (2021, p. 568). An AI that decides without provision of reasons—predictive forms of AI, tools generating credit scores, the risk of recidivism—preclude the possibility of gaining legitimacy through persuasion and consent (ibid.). "It seems no exaggeration to say," Herzog asserts, "that where algorithms make decisions, there is no 'common world' in the Arendtian sense, in which individuals jointly interpret the reality they encounter, exchange opinions, and come to judgments" (ibid., p. 570). Confronting an algorithmic system, "it does not make sense to try to convince it," giving rise to the prospect of "rule by nobody" that Arendt had feared (ibid.).

But what about language models that draw on party materials to make decisions supported by cogent and persuasive reasons? By virtue of taking into account both sides of an argument and providing reasons for a decision that seek to persuade us as fair, can these forms of AI not be said to engage a 'common world' or a kind of intersubjective process? Do they not anticipate what we—the parties, members of the community—will or will not likely consent to or accept through an appeal to reason? And do they not do so without rote application of general rules or standards? AI decisions might still be faulted for resting on opaque processes, for arriving at their outcome for statistical or arbitrary reasons; they might still entail a 'rule by nobody.' But the evolving capabilities canvassed in the experiments noted earlier complicate Herzog's criticism by providing a basis for questioning whether AI can approximate at least some aspects of what Arendt contemplated in relation to intersubjectivity in judgement.

Payman Tajalli (2021) foregrounds another aspect of Arendt's writing about morality and judgment to offer a further critical perspective on AI. Arendt's concern with persons who prove unwilling to judge, who fail to engage in the intersubjective processes outlined above, is related to a concern with what elsewhere she described as a failure to engage in thinking before deciding ethical issues Arendt (2003c). Arendt saw a close link between the "inability or refusal to think and the capacity of doing evil" (2003c, p. 180). Judgment does not depend on "a highly developed intelligence or sophistication in moral matters" but rather an ability to "be engaged in that silent dialogue between me and myself which… we usually call thinking" (2003a, p. 44–45). Making ethical decisions does not consist in "hold[ing] fast" to norms and standards because these "can be changed overnight"; more reliable, for Arendt, "will be the doubters and skeptics, not because scepticism is good or doubting wholesome, but because they are used to examine things and to make up their own minds" (ibid, p. 45). Tajalli argues that if thinking is essential to effective moral judgment as Arendt conceived it, AI could not be capable of it without being able to question the decisions it arrives at—or, "if need be, overrule the instructions coded in its own memory" (2021, p. 451). The challenge, however, is that thinking in this way cannot simply rely on "the execution of another set of codes or program"; thinking, he contends, "somehow needs to transcend the level of programmatic instruction execution" (ibid).

Tajalli concedes that if thinking were defined to consist of a "linear process of evaluation" and weighing of options, then AI may be capable of this (2021, p. 452)—and recent language models that incorporate 'reasoning' processes to guide their outputs might meet this criterion. But the question of what constitutes thinking is contested, with various arguments made about consciousness and subconsciousness, lived experience, and memory being integral parts of the process (Searle 1990; Schlagel 1999; Botică 2017). It may be that thinking is "not an all-or-nothing affair" but instead a "spectrum on which human beings, perhaps capable of doing maximal thinking, could be placed towards one end of the spectrum, and AI be moving towards this end" (Tajalli 2021, p. 452, citing Rapaport 1993, p. 18). It may be preferable to suggest that language models can engage in a form of thinking, different from but not necessarily inferior to the kind of thinking in which humans engage, and not necessarily inadequate to playing a role in legal or moral judgment.

The experiments canvassed earlier offer a basis for considering the possibility that AI can perform a function approximating moral judgment involving reflective thought—at least sufficient for some forms of legal judgment, including in challenging cases involving conflicts of rights or interests. This question would require a more detailed analysis of the cases tested in the experiments outlined above, the content of AI's output, and an interrogation of the underlying processes contributing to that output. What overlaying heuristics or refinements has a platform such as Claude or ChatGPT been programmed to include (in addition to the basic function of next-token completion) to help shape the final output, and how does this affect how the model takes a position on a moral question?

But apart from this more extensive inquiry, the experiments canvassed above demonstrate a model's ability to take a legal problem that contains a core moral or normative issue and offer a reasoned response—to take only one example: does a police demand for an Internet Protocol address engage a *reasonable* privacy interest? Evidence that something approximating thinking has taken place in arriving at a position on a normative question like this can be inferred from the fact that the model provides a justification for its decision (a reason why a thoughtful person should accept that an IP address does or does not attract a privacy interest). The justification segment in each of the cases in the experiments conducted by Unikowswki (2024a, 2024b), among others, tended to implicitly reject one party's view of the matter and favour the other by foregrounding as more deserving of general consent either the rationale the latter party offered or a better one the model produced on its own. Once again, this may not be thinking as Arendt conceived it, but it demonstrates a capability that earlier AI lacked and one that complicates the bare assertion that an AI cannot reach a decision that bears at least some of the hallmarks of reflective judgment.

3 Conclusion

The earlier part of this paper surveyed assumptions about AI's functional limitations that arose in a period in which forms of AI at issue tended to be narrow in scope (generating scores and predictions). AI involving language models that

draw on documents uploaded to the model are capable of making sophisticated legal judgments that challenge many of the earlier assumptions about AI's functional limits rendering it unsuitable for playing this role. Yet a number of powerful philosophical objections have remained. AI makes decisions on a computational or statistical rather than normative basis, and thus it cannot be said to decide on a rational or moral basis. A language model might avoid reliance on a simple form of if-then computation and might approximate the Aristotelian notion of *phronesis* by mimicking how humans decide what older situation a new moral conundrum is essentially like; but since a model does so in reliance on computation rather than intuition, it does so less reliably or effectively. And finally, in its reliance on computation and correlation, a language model cannot be said to engage in a true form of intersubjective and reflective thought, but only a simulation of this in reasons that are cogent and persuasive but arbitrary and hollow. These remain important differences between human and algorithmic judgment, and reasons to conclude the former cannot serve as a true substitute for, or be entirely replaced by, the latter.

What, then, should we make of AI's new capabilities—its ability to simulate normative reasoning, *phronesis*, and reflective judgment? One takeaway is that AI may be an effective aid to human judges or substitute for them in some contexts and under some conditions, including all-party consent. But this will depend in part on a shift in perceptions about the philosophical objections canvassed here in favour of a pragmatic view of automated judgment, at least in some contexts. To recall Eugene Volokh's position (2019), a pragmatic view would measure the acceptability of the AI judge in terms of the reliability, consistency, and quality of its judgments, and not whether they were arrived at on a normative or causal basis, whether they were based on a true capacity for *phronesis*, or whether they were the product of authentic, human intersubjective and reflective thought. That is to say, we might choose to rely on AI for judgment in some cases despite these reservations.

We may not have arrived at the point of fully meeting Volokh's judicial Turing test, given the lack of consistency and transparency in the best of the language models available at present. But the evolving capacity for judgment that language models have begun to demonstrate is moving us toward this point. In doing so, these better models do not refute the philosophical objections canvassed in this paper, but they point toward a moment in time when, in certain cases, they may not preclude relying on AI in judgment to a greater degree than was commonly assumed in the recent past.

Author Contributions R.D. is the sole author of the manuscript.

Funding The author received no funding for work on this paper.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Afrouzi AE (2024) Robots, Thurgood Martian, and the syntax monster: a new argument against AI judges. Can J Law Jurisprud 37(2):369

Arendt H (2003a) Personal responsibility under dictatorship. In: Kohn J (ed) Responsibility and judgment. Schocken Books, New York, pp 17–48

Arendt H (2003b) Some questions of moral philosophy. In: Kohn J (ed) Responsibility and judgment. Schocken Books, New York, pp 49–146

Arendt H (2003c) Thinking and moral considerations. In: Kohn J (ed) Responsibility and judgment. Schocken Books, New York, pp 159–189

Arrieta AB et al (2019) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. arXiv:1910.10045v2 [cs.AI]. Available at: https://arxiv.org/abs/1910.10045

Beatson J (2018) AI-supported adjudicators: should artificial intelligence have a role in tribunal adjudication? Can J Administrative Law Pract 31:307

Berman E (2018) A government of laws and not of machines. Boston Univ Law Rev 98(4):1277

Botică DA (2017) Artificial intelligence and the concept of human thinking. Business ethics and leadership from an Eastern European, transdisciplinary context. Springer, pp 87–94

Choi JH (2025) Large language models are unreliable legal interpreters. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5188865

Cofone IN (2021) AI and judicial decision-making. In: Martin-Bariteau F, Scassa T (eds) Artificial intelligence and the law in Canada. LexisNexis Canada, Toronto. Available at: https://ssrn.com/abstract=3733951

Coglianese C, Lehr D (2017) Regulating by robot: administrative decision making in the machine-learning era. Georget Law J 105(5):1147

Crootof R (2019) Cyborg justice' and the risk of technological-legal lock-in. Columbia Law Rev Forum 119:233

Daly P (2023) Artificial administration: administrative law, administrative justice and accountability in the age of machines. Aust J Adm Law Pract 30(2):95. Available at: https://ssrn.com/abstract=4434238

Diab R (2026) The evolving role of AI in legal judgment. Law Innovation and Technology. Available at: https://doi.org/10.1080/17579961.2026.2633688

Eisikovits N, Feldman D (2022) AI ethics beyond principles: strengthening the life-world perspective. Moral Philos Polit 9(2):181–199. Available at: https://doi.org/10.1515/mopp-2021-0026

Frost N (2024) The impoverished publicness of algorithmic decision making. Oxf J Legal Stud 44(4):780

Gandall K, Kieffaber J, McLaren K (2025) We built Judge.AI and you should buy it. Available at: https://ssrn.com/abstract=5115184

Garsten B (2007) The elusiveness of Arendtian judgment. Soc Res 74(4):1071–1108 Available at:. www.jstor.org/stable/40972041

Gouge A (2021) Administrative law, artificial intelligence, and procedural rights. Windsor Rev Legal Social Issues 42:17

Grimmelmann J, Sobel BLW, Stein D (2026) Generative misinterpretation. Harv J Legislation 63(1):229–308

Groff R, Symons J (2024) Is AI capable of Aristotelian full moral virtue? The rational power of phronesis, machine learning and regularity. In: Bauer WA, Marmodoro A (eds) Artificial dispositions: investigating ethical and metaphysical issues. Bloomsbury, London, pp. 219–232.

Herzog L (2021) Old facts, new beginnings: thinking with Arendt about algorithmic decision-making. Rev Politics 83(4):555–577

Hildebrandt M (2008) Legal and technological normativity: more (and less) than twin sisters. Techné 12(3):169

Kahneman D, Sibony O, Sunstein CR (2021) Noise. A flaw in human judgment. Little, Brown Spark, New York

Kieffaber J, Gandall K, Foster S, McLaren K (2025) LLMs are bad judges. So use our classifier instead. SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5331811

Maruthi S et al (2022) Language model interpretability – explainable AI methods. Aust J Mach Learn Res Appl 2(2). Available at: https://sydneyacademics.com/index.php/ajmlra/article/view/19

Nietzsche F (1990) Beyond good and evil. Vintage Books, New York

Posner EA, Saran S (2025) Judge AI: assessing large language models in judicial decision-making. University of Chicago Coase-Sandor Institute for Law & Economics Research Paper No 25–03. Available at: https://ssrn.com/abstract=5098708
Rapaport WJ (1993) Because mere calculating isn't thinking. Mind Mach 3:11–20
Raso J (2021) AI and administrative law. In: Martin-Bariteau F, Scassa T (eds) Artificial intelligence and the law in Canada. LexisNexis Canada, Toronto, pp. 163–184.
Re RM, Solow-Niederman A (2019) Developing artificially intelligent justice. Stanf Technol Law Rev 22(2):242
Schlagel RH (1999) Why not artificial consciousness or thought? Mind Mach 9:3–28
Searle JR (1990) Is the brain's mind a computer program? Sci Am 262(1):25–31
Shay LA, Hartzog W, Nelson J, Conti G (2016) Do robots dream of electric laws? An experiment in the law as algorithm. In: Calo R, Froomkin AM, Kerr I (eds) Robot law. Edward Elgar, Cheltenham, p 274
Sourdin T, Cornes R (2018) Do judges need to be human? The implications of technology for responsive judging. In: Sourdin T, Zariski A (eds) The responsive judge: international perspectives. Springer, New York, pp 87–119
Sullens JP (2021) Artificial phronesis: what it is and what it is not. In: Ratti E, Stapleford TA (eds) Science, technology, and virtues: contemporary perspectives. Oxford University Press, Oxford. Available at: https://doi.org/10.1093/oso/9780190081713.003.0008
Sunstein CR (2001) Of artificial intelligence and legal reasoning. University of Chicago Law School Roundtable 8:29. Available at: https://chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=12376&context=journal_articles
Surden H (2019) Artificial intelligence and law: an overview. Ga State Univ Law Rev 35(4):1305
Susskind R (2019) Online courts and the future of justice. Oxford University Press, Oxford
Tajalli P (2021) AI ethics and the banality of evil. Ethics and Information Technology 23:447–454. Available at: https://doi.org/10.1007/s10676-021-09587-x
Tasioulas J (2023) The rule of algorithm and the rule of law. Vienna Lectures on Legal Philosophy. Available at: https://ssrn.com/abstract=4319969
Tömmel T, Passerin d'Entreves M (2025) Hannah Arendt. In: Zalta EN, Nodelman U (eds) The Stanford Encyclopedia of Philosophy, Spring 2025 edn. Available at: https://plato.stanford.edu/archives/spr2025/entries/arendt/
Tsai CH, Ku H (2025) Why AI may undermine phronesis and what to do about it. AI Ethics 5:3079–3086. Available at: https://doi.org/10.1007/s43681-024-00617-0
Unikowsky A (2024a) In AI we trust. Adam's Legal Newsletter. Available at: https://adamunikowsky.substack.com/p/in-ai-we-trust
Unikowsky A (2024b) In AI we trust, part II. Adam's Legal Newsletter. Available at: https://adamunikowsky.substack.com/p/in-ai-we-trust-part-ii
Volokh E (2019) Chief justice robots. Duke law J 68(6):1135
Williams R (2022) Rethinking administrative law for algorithmic decision making. Oxf J Legal Stud 42:468
Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, Wang S, Yin D, Du M (2024) Explainability for large language models: a survey. ACM Trans Intell Syst Technol 15(2). Available at: https://doi.org/10.1145/3639372