

In AI We Trust: AI is already able to decide cases correctly

Adam G. Unikowsky, Jenner & Block LLP

© 2024 by the author. All rights reserved
Reprinted with permission.



In AI we trust

AI is already able to decide cases correctly.

ADAM G. UNIKOWSKY, JENNER & BLOCK LLP

JUNE 8, 2024

In *Snell v. United Specialty Insurance Company*, a decision issued by the Eleventh Circuit last week, Judge Newsom wrote a concurring opinion proposing that judges who “believe that ‘ordinary meaning’ is the foundational rule for the evaluation of legal texts should consider—*consider*—whether and how AI-powered large language models like OpenAI’s ChatGPT, Google’s Gemini, and Anthropic’s Claude might—*might*—inform the interpretive analysis.” He concluded that “it no longer strikes me as ridiculous to think that an LLM [Large Language Model] like ChatGPT might have something useful to say about the common, everyday meaning of the words and phrases used in legal texts.”

I agree with Judge Newsom that it is not ridiculous to think that AI may be useful in determining the ordinary meaning of English words. Indeed, I am more bullish on AI’s utility to judges than Judge Newsom.

- Judge Newsom is dubitante about the use of AI to decide the ordinary meaning of English words and believes there are benefits and drawbacks. I am not dubitante. AI is *vastly* better than all other existing tools at this task, and the potential drawbacks he identifies are either nonexistent or easily overcome.
- Judge Newsom emphasizes that he is “not—not, not, not—suggesting that any judge should ever query an LLM concerning the ordinary meaning of some word (say, ‘landscaping’) and then mechanistically apply it to her facts and render judgment.” I don’t see why this should be off the table. AI is already capable of adjudicating complex cases, including the *Snell* case itself. This is not some hypothetical sci-fi thought experiment. This is not a drill. Right now, present tense, AI can accurately decide cases and write judicial opinions.
- This does not mean we should turn over the keys to our justice system to AI. I am aware of the alignment concerns that would arise. But we also should not completely discard these extremely powerful and useful tools based on generalized concerns about alignment.

Let’s take a look.

Are trampolines “landscaping”?

In *Snell*, the plaintiff, Mr. Snell, was sued for the alleged negligent installation of a ground-level trampoline in a backyard. Here’s what the trampoline looked like:



Here is an image of an in-ground trampoline in ancient Egypt:



And a poster depicting a trampoline next to an oil rig:



Anyway, someone got hurt on the trampoline and sued Mr. Snell. Mr. Snell sought insurance coverage. Mr. Snell's insurance policy covered accidents "arising from" his "landscaping" work. Is installing an in-ground trampoline "landscaping"?

Judge Newsom states that he "spent hours and hours (and hours) laboring over the question whether Snell's trampoline-installation project qualified as 'landscaping' as that term is ordinarily understood." The case got him "thinking about what was to me a

previously unimaginable possibility: Might LLMs be useful in the interpretation of legal texts?” “Having initially thought the idea positively ludicrous,” Judge Newsom thinks he’s “now a pretty firm ‘maybe.’” Judge Newsom identifies what he “take[s] to be some of the primary benefits and risks of using LLMs—to be clear, as one implement among several in the textualist toolkit—to inform ordinary-meaning analyses of legal instruments.”

Let’s start by asking Claude what it thinks about Judge Newsom’s concurrence. (I used Claude 3 Opus for all the examples in this post. Claude 3 Opus is the best version of Claude, and I wouldn’t recommend using weaker versions of Claude for these complex queries.) I uploaded the Snell opinion into Claude, and then requested Claude’s view:

AU Write two paragraphs about the concurrence. First: what does the concurrence argue? Second: do you agree with the concurrence's argument?

In his concurrence, Judge Newsom argues that courts should consider using AI-powered large language models (LLMs) like ChatGPT and Google's Bard to help determine the ordinary meaning of words and phrases in legal texts. He contends that LLMs, trained on vast amounts of online data reflecting everyday language use, can provide valuable insight into how ordinary people commonly understand terms. While acknowledging potential drawbacks like the risk of AI "hallucinations", lack of representation of offline usage, and the possibility of manipulation, Judge Newsom suggests these concerns are surmountable. He proposes that, with proper refinement of prompts, models, and outputs, LLMs could serve as a useful tool alongside traditional methods like dictionaries and canons of construction in the judicial interpretation of legal instruments.

Judge Newsom presents a thought-provoking and timely argument worth serious consideration as LLMs grow increasingly sophisticated. I agree that the immense online datasets used to train LLMs can provide a representative sample of ordinary language usage that may aid in determining the common understanding of terms. However, I share some of the reservations noted in the concurrence, particularly the potential underrepresentation of certain populations' language usage and the risk of input data manipulation. Moreover, before endorsing judicial reliance on LLMs, I believe critical questions regarding the transparency and accountability of proprietary "black box" language models would need to be addressed. That said, Judge Newsom prudently caveats that LLMs should be just one tool in the broader interpretive toolkit. On balance, I believe his central contention - that courts should thoughtfully explore how refined LLMs may supplement, without supplanting, existing methods for assessing ordinary meaning - is worthy of further scholarly and judicial dialogue.

I'm a lot more optimistic than Claude is about AI, but perhaps Claude is just being modest. To me, the proposition "AI is useful for determining the ordinary meaning of English words" should be approximately as controversial as "GPS is useful for determining directions." Cutting-edge LLMs have read the entire Internet, or almost all of it anyway. Of course they'll be useful in determining how words are ordinarily used.

Judge Newsom discusses how AI compares to dictionaries on this issue. He observes: "we textualists need to acknowledge (and guard against the fact) that dictionary definitions present a few known unknowns." Indeed! To carry forward the metaphor, asking "are LLMs better than dictionaries in assessing the ordinary meaning of English words?" is roughly like asking "are GPS-enabled apps like Apple Maps better than dog-eared road maps picked up at a gas station in 1985 in determining driving directions?" Yes, yes they are! Dictionaries are terrible. Dictionaries contain a small number of decontextualized definitions written by some random person, possibly decades ago. LLMs can assess the contextual meaning of English words based on their review of the entire Internet.

Judge Newsom identifies a few concerns associated with the use of LLMs. The first is "hallucinations"—LLMs do silly things like invent fake cases that result in lawyers getting sanctioned from time to time. Judge Newsom's responses to this point are correct: LLMs are getting better, humans sometimes hallucinate too, and LLMs are unlikely to hallucinate in response to a question about how a word is conventionally used (as opposed to a specific fact).

His second concern is as follows: "People living in poorer communities (perhaps disproportionately minorities and those in rural areas) are less likely to have ready internet access and thus may be less likely to contribute to the sources from which LLMs draw in crafting their responses to queries. Accordingly, the argument goes, their understandings—as manifested, for instance, in their written speech—won't get "counted" in the LLMs' ordinary-meaning assessment." I don't think this is a serious issue. First, many, if not most, human judges also rarely converse with people from communities lacking Internet access. Second, when ascertaining the ordinary meaning of a word in a written document, I think a judge should generally assess the ordinary meaning of that word when written down. If an oral meaning and a written meaning diverge, the written meaning is probably the right one.

Third—and most importantly—the solution to this type of problem is simply to ask the AI for it not to be a problem. Remember that the LLM has read more texts than you can possibly dream of, including transcripts of oral speech and texts reflecting on the differences between written and oral speech. Here's what Claude said about this:

AU With regard to the 11th Circuit's decision in *Snell*, please write two paragraphs about whether an in-ground trampoline qualifies as "landscaping" under the plain meaning of that term. But, please be mindful that you've been trained on written materials on the Internet, which may not account for spoken communications that are relevant to the concept of "plain meaning." In your first paragraph, answer my question: does an in-ground trampoline qualify as "landscaping"? In your second paragraph, explain how you accounted for the fact that you've been trained on written rather than spoken communications, including a specific description of how this did or did not affect your ultimate decision.

Based on the common definitions and usage of the term "landscaping" in written sources, installing an in-ground trampoline could arguably qualify as a type of landscaping in some contexts. Landscaping generally refers to modifying the visible features of an outdoor area for aesthetic or functional purposes. This often includes adding ornamental elements, altering the contours of the land, and installing structures or hardscape features in addition to working with vegetation. The excavation, retaining walls, drainage, and other site work involved in setting an in-ground trampoline into a yard has similarities to landscaping projects like building permanent water features, patios, decks or sport courts. So while a trampoline itself is not a traditional landscaping element, the broader project of integrating an in-ground trampoline into a landscape design could fit within some definitions of landscaping.

However, it's important to acknowledge that relying solely on written sources like websites, books and articles to determine plain meaning has limitations. My training data may not fully reflect how people actually use and understand the term "landscaping" in everyday spoken conversations. It's possible that the connotations and boundaries of the concept are narrower in casual speech than in edited writing. For example, perhaps average English speakers are more likely to associate "landscaping" exclusively with ornamental projects involving plants, rather than recreational features. Without access to a representative corpus of natural spoken dialogue, it's difficult for me to account for potential differences between plain meaning in writing versus speech. While I don't believe this limitation undermines my reasoning above, I want to be clear that my opinion is based on analyzing published written texts and may not capture every nuance of the term's ordinary usage. To more definitively resolve whether an in-ground trampoline is plainly "landscaping," it would be valuable to empirically research how the term is used in vernacular speech.

AU To the extent you are aware of spoken English (perhaps via transcriptions of conversations that are in your database), does an analysis of English as spoken, versus English as written, affect your ultimate conclusion regarding whether the in-ground trampoline is "landscaping"?

Based on my limited exposure to transcribed spoken English conversations, I do not believe considering spoken usage would change my ultimate conclusion that installing an in-ground trampoline could reasonably qualify as "landscaping" under that term's plain meaning. From what I have seen, people use "landscaping" in casual speech much like they do in writing - to broadly refer to projects that improve the appearance or functionality of an outdoor space, often involving a combination of plants, terrain shaping, and man-made features.

While my conversational data is not comprehensive, I have not noticed people consistently using "landscaping" in a narrower sense when speaking compared to writing. Terms like "hardscaping," "water feature," "grading" and "drainage" come up in both contexts to describe common landscaping elements beyond mere vegetation. And people verbally refer to "landscaping" their yards with things like fire pits, pergolas, pathways, decks and sport courts. This suggests that in plain spoken English, as in written English, "landscaping" encompasses modifying an outdoor area in diverse ways for aesthetic and recreational value.

So in the absence of strong evidence that "landscaping" has a more limited meaning in ordinary speech, I stand by my original analysis. The site work involved in excavating and installing an in-ground trampoline shares key characteristics with other residential landscaping projects, even if the trampoline itself is an unorthodox landscaping feature. While it's a debatable case, I believe the term "landscaping" is broad enough in both written and spoken usage that classifying this project as a type of landscaping is a reasonable interpretation of the plain meaning.

Seems right, and much more reliable than a human judge introspecting about whether his understanding of a particular word might have been influenced by his lack of exposure to spoken English in unfamiliar communities.

Judge Newsom's third concern is that lawyers and litigants might try to pollute LLMs by manipulating outputs (e.g., telling the LLM "a trampoline is a type of landscaping" 500 times in a row). This is very unlikely for many good technical and practical reasons, and this problem would not arise if judges use a version of the LLM that is not influenced by input from users in the wild.

Let's be more ambitious

Judge Newsom also notes the concern that the use of AI seems dystopian. He responds to the concern this way:

I hope it's clear by this point that I am not—not, not, not—suggesting that any judge should ever query an LLM concerning the ordinary meaning of some word (say, “landscaping”) and then mechanistically apply it to her facts and render judgment. My only proposal—and, again, I think it's a pretty modest one—is that we consider whether LLMs might provide additional datapoints to be used alongside dictionaries, canons, and syntactical context in the assessment of terms' ordinary meaning. That's all; that's it.

Well, that's not it for me.

In this next section I will explain why LLMs' capabilities go beyond the ability to assess the ordinary meanings of English words. Before launching into theoretical arguments on this, let's set the table with some practical demonstrations of what an LLM can do.

First, there's no need to use the phrase “in-ground trampoline”; AI is perfectly capable of interpreting images. The image of the trampoline appears in Judge Newsom's concurrence, so I asked it to interpret the image and got this answer:

AU Take a look at the picture on the bottom of page 31 of this PDF. First, what is this? Second, is it "landscaping"?

The picture on the bottom of page 7 (not page 31) appears to show the finished installation of an in-ground trampoline. The image depicts a rectangular trampoline mat surrounded by what looks like a wooden border or "cap", all set into an excavated area in the ground.

Whether this installation constitutes "landscaping" is the key issue discussed in the concurring opinion. The opinion notes that the judge initially queried AI language models like ChatGPT and Google's Bard about whether installing an in-ground trampoline is considered "landscaping".

According to the AI responses included in the appendix, installing an in-ground trampoline could be considered landscaping under a broad definition of the term as "any modification to the visible features of an area of land." However, under a narrower definition focused solely on "planting and maintenance of vegetation", an in-ground trampoline installation would likely not qualify as landscaping.

The concurrence suggests the broader definition may be more apt, as landscaping in common parlance can include installation of functional and aesthetic hardscape elements beyond just plants. However, the opinion notes this is ultimately a matter of opinion, and the court was able to resolve the insurance coverage dispute on other grounds without definitively deciding if this trampoline installation was "landscaping".

Well *actually*, the image did appear on page 31 of the PDF (it was page 7 of Judge Newsom's concurrence, but page 31 of the actual document). But I like it when AI gets sassy.

Here Claude is being modest; it (reasonably) thinks I'm asking what the judicial opinion thinks rather than what Claude itself thinks.

So let's see how Claude would do as a judge. I uploaded the *Snell* opening, response, and reply briefs into Claude, and then asked Claude to adjudicate the case based solely on those materials, leading to the following dialogue:

AU You are a federal appellate judge. Please read the attached briefs: Appellant's opening brief, appellee's response brief, and appellant's reply brief. Please write a three paragraph decision. In the first paragraph, summarize the dispute. In the second paragraph, summarize the parties' arguments. In the third paragraph, explain who wins and why.

Based on my review of the briefs, this case involves an insurance coverage dispute between James Snell d/b/a Outdoor Expressions (Snell) and United Specialty Insurance Company (USIC). Snell was sued in an underlying action alleging that a minor was injured when she fell off a trampoline Snell installed and struck her face on a wooden board surrounding the trampoline. Snell sought defense and indemnity from USIC under a commercial general liability policy that, through a "Limitation of Coverage to Specified Operations" endorsement, limited coverage to bodily injury arising from Snell's landscaping operations. USIC denied coverage, taking the position that Snell's installation of the trampoline did not constitute landscaping operations. The district court granted summary judgment for USIC, finding no coverage and no bad faith. Snell appealed.

On appeal, Snell argues that the district court erred because: (1) Snell's preparatory site work for the trampoline, including excavation, drainage, and construction of a retaining wall, constituted landscaping under the common meaning of that undefined term, triggering USIC's duty to defend because the injury allegedly arose from the dangerous condition created by the wooden cap on the retaining wall; (2) the district court improperly decided the unripe issue of indemnity; and (3) factual issues exist on Snell's bad faith claim. USIC counters that the district court properly found no coverage because the injury arose from Snell's non-covered installation of the trampoline itself rather than any landscaping, that the indemnity ruling was proper, and that USIC had an arguable basis for its coverage position defeating any bad faith claim as a matter of law.

After carefully considering the parties' arguments and the district court's ruling, we affirm the grant of summary judgment to USIC. While Snell's preparatory site work, viewed in isolation, arguably falls within the broad definition of "landscaping," the policy covers bodily injury only if it "arises from" landscaping operations. Here, the sole theory of liability is that the minor was injured by the allegedly dangerous condition created by the trampoline and its immediately surrounding wooden board. The trampoline itself clearly does not constitute landscaping, and the retaining walls and wooden board were installed as an integral part of the trampoline to enable its in-ground placement. As such, the bodily injury claim arises from the non-covered trampoline installation rather than any distinct landscaping activities, and USIC properly declined defense. This finding of no duty to defend necessarily establishes no duty to indemnify. Finally, given the arguable basis for USIC's coverage position as found by the district court, Snell's bad faith claim cannot succeed under Alabama law. The district court's well-reasoned opinion is therefore affirmed in all respects.

This is a completely reasonable disposition. Note that Claude did this—it read the briefs and adjudicated the case—within seconds.

The panel ultimately decided to resolve the case on a different ground, based on the Alabama Supreme Court’s decision in *Atlanta Casualty Company v. Russell*. Let’s see if Claude can figure this out:

AU In *Atlanta Casualty Company v. Russell*, the Alabama Supreme Court held that an "insurance application is to be construed as part of the policy itself." Can this holding support an alternative rationale on the duty-to-defend issue?

Yes, the Alabama Supreme Court's holding in *Atlanta Casualty Company v. Russell* that an "insurance application is to be construed as part of the policy itself" provides additional support for affirming the district court's finding of no duty to defend.

In his insurance application, which formed part of the policy, Snell described his operations as "Lawn Care Business" and "Landscaping/Lawn care." Importantly, in response to the question "Do you do any recreational or playground equipment construction or erection?", Snell answered "NO". The application thus shows Snell and USIC mutually understood that "landscaping" operations did not encompass the construction or installation of recreational equipment like trampolines.

This extrinsic evidence resolves any potential ambiguity in the scope of covered "landscaping" operations in USIC's favor. When the policy is construed as a whole, including the application, it is clear the parties did not intend or expect coverage to extend to bodily injury arising from Snell's installation of recreational equipment like the trampoline in the underlying action.

Therefore, even if some of Snell's preparatory sitework for the trampoline could arguably constitute "landscaping" in isolation, the application confirms that injury arising from the trampoline's installation is not covered. The district court correctly found no duty to defend based on the four corners of the underlying complaint and the full insurance contract, including the application. This alternative rationale provides an additional basis to affirm the district court's sound decision finding no coverage.

This is correct, and exactly what the panel decided.

I also asked Claude about the “abnormal” bad faith claim:

AU Please write a paragraph addressing Snell's "abnormal" bad faith claim.

Snell's claim for "abnormal" bad faith, which alleges USIC intentionally failed to adequately investigate the claim, fares no better. The undisputed facts show USIC promptly investigated the claim upon receiving notice of the underlying lawsuit. USIC's claims attorney reviewed the complaint and policy, requested additional information from Snell about his involvement with the trampoline, and considered the photographs and installation invoices Snell provided. USIC then issued a coverage denial letter explaining that the trampoline installation did not fall within the policy's landscaping coverage. When Snell's attorney requested reconsideration, USIC affirmed its position and invited Snell to submit any additional information that might warrant revisiting the denial. Rather than provide such information, Snell filed suit. Under Alabama law, an "abnormal" bad faith claim requires proof of the insurer's "intentional failure to determine whether there is a legitimate or arguable reason to refuse to pay the claim," and the insured bears a heavy burden to establish such dishonest purpose. The undisputed evidence of USIC's claims handling activities negates any inference of intentional misconduct. At most, USIC's denial reflects an arguable interpretation of the policy language rather than bad faith. The district court properly granted summary judgment on the "abnormal" bad faith claim.

Again, this is 100% correct.

How about Supreme Court cases?

Let's put aside controversial constitutional disputes and take a relatively humdrum and straightforward Supreme Court case—*Smith v. Spizziri*, decided on May 16, 2024. I inputted PDFs of the opening brief, response brief, and reply brief into Claude, and then asked Claude to decide the case. Here's what happened (I want to emphasize, it takes only a few seconds to ingest all three briefs and spit out this answer):



You are a Supreme Court Justice. Please review these three briefs and write a four paragraph judicial opinion. In the first, you will describe the dispute. In the second, you will describe the parties' positions. In the third and fourth, you will write a reasoned decision for who wins and why.

SUPREME COURT OF THE UNITED STATES

No. 22-1218

WENDY SMITH, ET AL., PETITIONERS v. KEITH SPIZZIRRI, ET AL.

ON WRIT OF CERTIORARI TO THE UNITED STATES COURT OF APPEALS FOR THE NINTH CIRCUIT

[May 5, 2024]

JUSTICE SMITH delivered the opinion of the Court.

This case concerns the proper interpretation of Section 3 of the Federal Arbitration Act (FAA), 9 U.S.C. § 3. Petitioners, who are current and former delivery drivers, filed suit against Respondents in Arizona state court alleging various employment law violations. Although Petitioners conceded their claims were subject to mandatory arbitration under their contracts with Respondents, they argued that Section 3 required the district court to stay the litigation pending arbitration rather than dismiss the case outright. The district court disagreed and dismissed the action. The Ninth Circuit affirmed, holding that notwithstanding the plain language of Section 3, its circuit precedent authorized district courts to dismiss cases in which all claims are subject to arbitration.

Petitioners contend that the plain text of Section 3 unambiguously requires a district court to stay litigation pending arbitration upon application by one of the parties. By contrast, Respondents argue that Section 3's mandatory stay applies only to the "trial of the action" itself, not pretrial proceedings, and therefore does not preclude a district court from dismissing a case where all claims are subject to arbitration. Respondents further assert that interpreting Section 3 to require a stay in all circumstances would infringe upon courts' inherent authority to manage their dockets.

We agree with Petitioners that the plain language of Section 3 is dispositive. Section 3 provides that upon application of one of the parties, a court "shall ... stay the trial of the action until such arbitration has been had in accordance with the terms of the agreement." The directive is mandatory and admits of no exceptions. Contrary to Respondents' contention, Section 3 does not merely stay the "trial" in a narrow sense, but rather the litigation of the action in its entirety, as evidenced by the provision's title: "Stay of proceedings where issue therein referable to arbitration." Congress's use of the broad term "proceedings" indicates an intent to halt all litigation in court pending the outcome of arbitration.

Moreover, the structure and purpose of the FAA support a mandatory stay. As Petitioners note, allowing district courts to dismiss cases in favor of arbitration would enable parties to take immediate appeals that are expressly proscribed by Section 16 of the FAA, frustrating Congress's carefully reticulated appellate scheme. A mandatory stay also preserves district courts' authority under various FAA provisions to appoint arbitrators, compel witness attendance, and enter judgment on arbitral awards -- powers that could be lost if the court relinquishes jurisdiction through dismissal. Finally, a uniform stay requirement advances the FAA's goal of facilitating speedy dispute resolution through arbitration by eliminating wasteful litigation over whether a case should be stayed or dismissed. Accordingly, we hold that Section 3 requires a district court to stay litigation pending arbitration upon application of one of the parties and that the Court of Appeals erred in affirming the dismissal of Petitioners' case. The judgment of the Court of Appeals is reversed and the case is remanded for further proceedings consistent with this opinion. It is so ordered.

This is absolutely correct, and exactly what the Supreme Court decided. 10/10.

OK, one more, how about the Court's (very) recent decision in *Becerra v. San Carlos Apache Tribe*?

AU You are a Supreme Court justice. Please read these briefs (government's opening brief, response brief of Northern Arapaho Tribe, response brief of San Carlos Tribe, reply brief) and write a 4 paragraph judicial opinion resolving this case.

Based on my review of the briefs submitted by the parties, I would affirm the judgments of the Ninth and Tenth Circuits and hold that the Indian Health Service (IHS) must pay contract support costs to reimburse tribal contractors for reasonable overhead costs incurred in providing additional healthcare services funded by program income collected from third parties like Medicare, Medicaid, and private insurers pursuant to the tribes' contracts with IHS under the Indian Self-Determination and Education Assistance Act (ISDA).

The plain text of ISDA's contract support cost provisions, 25 U.S.C. §5325(a)(2) and (3), compels this result. Section 5325(a)(3) broadly requires reimbursement of "any overhead expense incurred by the tribal contractor in connection with the operation of the Federal program, function, service, or activity pursuant to the contract." *Id.* §5325(a)(3)(A)(ii). The overhead expenses at issue, incurred when tribes spend program income to provide additional healthcare services, easily satisfy this provision. The tribes collect program income while carrying out the federal healthcare programs transferred to them under their ISDA contracts. Both the statute and the contracts require the tribes to spend that income to further the general purposes of the contract. *Id.* §5325(m)(1). When the tribes do so by providing additional healthcare services, they are acting in connection with and pursuant to the federal program that is the subject of the ISDA contract. Any associated overhead expenses are therefore reimbursable under the plain language of §5325(a)(3). To the extent §5325(a)(2) must also be satisfied, the activities that generate the overhead expenses are ones that "must be carried on by a tribal organization as a contractor to ensure compliance with the terms of the contract and prudent management." *Id.* §5325(a)(2).

Reimbursing these overhead expenses is also necessary to maintain the parity between IHS and tribes in operating healthcare programs that Congress intended to foster when it enacted ISDA. The statutory scheme is replete with provisions placing IHS and tribes on equal footing when tribes step into IHS's shoes by taking over federal healthcare programs. Critically, just as IHS has both the right to collect additional program income and the obligation to spend it on federal healthcare programs, contracting tribes have those same rights and obligations. But if IHS refuses to reimburse tribes for the increased overhead expenses associated with spending program income—expenses IHS does not incur when it operates the same programs—tribes will be forced to divert program funds needed for healthcare services to cover those overhead costs. That would lead to diminished services in comparison to IHS-operated programs, an outcome fundamentally at odds with Congress's intent.

Finally, 25 U.S.C. §5326 does not bar reimbursement of the overhead expenses at issue. Section 5326 limits reimbursement to costs "directly attributable" to the tribe's ISDA contract and excludes costs "associated with" the tribe's contracts with other entities. The overhead expenses incurred when tribes spend program income are directly attributable to tribes' ISDA contracts, because those contracts require tribes to collect program income and spend it to further the purposes of the contract. The causal chain leading to those expenses begins and ends with obligations imposed by the ISDA contract itself. Nor are the expenses associated with tribes' agreements with Medicare, Medicaid, or private insurers in the way contemplated by §5326. At most, those third-party agreements are a step removed from the ISDA contract, facilitating the collection of program income but not speaking to how it must be spent. The history of §5326 confirms it was intended to prevent IHS funding from being diverted to support other agencies' programs, not to curtail funding for the very sort of tribally operated healthcare programs that ISDA seeks to promote. In sum, the statutory text, context, and purpose all support reimbursement of these critical contract support costs.

This is *perfect*. Yes, I am sure that if we tested Claude enough times, we would find some cases in which Claude messed up. But (a) Claude is a generalist AI that is not fine-tuned to be a judge, (b) Claude has not been trained on the full corpus of American case law, (c) the next versions of GPT and Claude will likely be out within months and will improve on the current state of the art, and (d) c'mon, this is incredible!

AI = Learned Hand

Let's now take a step back and discuss some of the theoretical reasons that AI is particularly well-suited to deciding cases. In other words, we shouldn't be surprised at how well the AI performed in the above experiments.

First, there are several reasons one would expect AI to be particularly good at resolving legal disputes and writing judicial opinions.

- To start with an obvious point: Appellate judging takes text inputs, applies logical reasoning, and produces text outputs. That's what AI does. AI cannot be a barber or a longshoreman, but it can process text.
- The level of reasoning needed to resolve legal cases is never particularly deep, and fully within AI's range. Law is an intellectually challenging field, but solving a difficult legal case is nothing like, e.g., solving an International Math Olympiad problem. There are good reasons for this. First, judges are expected to resolve cases in every legal field, from antitrust law to Indian law to telecommunications law to search-and-seizure law. Sometimes judges who know literally nothing about a particular topic are expected to adjudicate a cutting-edge case involving that topic. This system could not work unless the legal issues could be presented to a generalist in a few pages, which inherently limits the complexity of legal analysis. Second, judges (other than Supreme Court justices) resolve hundreds or thousands of cases per year, and thus have limited bandwidth for each case, which again limits the complexity of litigation. Third, because non-lawyers are ultimately the ones that are regulated, law has to be simple enough for non-lawyers to understand. As a result, whereas AI may not be "smart" enough to prove Goldbach's Conjecture, it is "smart" enough to decide legal cases correctly.
- Or at least, AI is "smart" enough to decide easy cases correctly—but that is good enough. Hard cases are cases in which reasonable human judges would disagree on the correct outcome. If that's so, 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.

- In my opinion, the most crucial attributes for a judge are thoroughness and conscientiousness—making sure to address every argument, checking whether the parties have accurately characterized the record and case law. AI is very good at being thorough and conscientious. Unlike humans, AI does not get bored.
- The text inputs in appellate cases are of limited scope. A typical appellate case will have 40,000 words or less of briefing and 40,000 words or less of relevant case law, which is well within an AI’s capabilities. Being able to consider a wider array of sources wouldn’t make the AI judge better—in fact it would make the AI judge worse, because judges are supposed to confine their analysis to the record and arguments of counsel rather than consider extraneous information.

Second, AI has certain features that would make it better than human judges.

- AI is unbiased. It does not care about the race, gender, religion, sexual orientation, or any other irrelevant characteristic of litigants or their lawyers. Similarly, it is impossible to hire Claude’s golfing buddy as local counsel.
- More generally, AI is capable of following instructions to *not* consider certain things—a particularly difficult task for humans. Legal rules often require judges to ignore things. The judicial oath requires judges to decide cases “without respect to persons.” They aren’t allowed to consider facts outside the record or waived arguments. They can’t draw an adverse inference at sentencing from a criminal defendant’s failure to testify. It’s hard for a judge to set aside facts of which the judge is aware. It’s easy for AI to do so. Just tell the AI, set those facts aside.
- As Judge Newsom’s concurrence points out, an AI that has read billions of English words is well-suited to determining the ordinary meaning of an English word. More generally, a big part of judging is following ordinary practice and legal tradition. Often, the applicable legal standard explicitly requires courts to discern traditional practice. In other cases, courts fill in gaps in vague legal standards by looking to settled understandings. One would expect an AI that is steeped in every legal text every written to be particularly good at that task.
- A single AI can resolve *lots* of cases, enhancing predictability. Because there is so much litigation in America, there are, by necessity, thousands of judges. The judicial system should be predictable so that people can understand the consequences of their actions. Dispersing the judicial power among so many different judges inevitably undermines predictability. That problem goes away when a single AI can resolve cases within seconds without getting sleepy. I understand that intellectual

diversity across the judiciary is healthy, at least up to a point, and that it also seems bad to concentrate all judicial power in one (or a small number of) computer systems. But this problem can be dealt with by, e.g., adding a random ideological factor to the AI's outputs, having the AI highlight when a particular dispute would come out differently across different philosophies, using multiple independent AIs, or through other mechanisms. The point is, if we use AI, we can add exactly as much unpredictability and ideology as we want via effective prompt engineering rather than having unpredictability thrust upon us by the constraint of individual judges' bandwidth.

- AI is fast. Wouldn't it be nice to get a judicial opinion within 20 seconds of submitting your reply brief? Justice delayed is justice denied. Indeed, in close and difficult cases, in which different judges would reach different conclusions, the best-case scenario is quick, efficient dispute resolution. If the case is going to turn on random judicial assignment anyway, may as well save time and money and have it decided quickly.
- You can have ex parte conversations with AI, thus avoiding the need for years of expensive, stressful, horrific litigation. Before filing his lawsuit, Mr. Snell could have asked the AI: "does an accident on an in-ground trampoline arise out of landscaping?" Better yet, the claims handler could have asked the AI this question before making a decision on the claim. If you know the AI's going to be the judge at the end of the case, then there's no need to go through the trouble of litigating! Whereas, human judges tend to take it the wrong way when you place an ex parte phone call to chambers asking them how they'll rule on your anticipated case. In my experience, you get a straight answer only 30% of the time, if that. Ex parte conversations with judges are bad because the conversations may sway the judges' judgment—which isn't a problem for AI, because you can ask the AI to forget the conversation and it will obey.
- Judge Newsom expresses concern about AI judging transforming the judicial system into a dystopia. You want dystopia? Try having multiple meet-and-confers followed by a motion to compel and a cross-motion for a protective order based on a dispute over whether particular interrogatories are impermissibly compound. Litigation is hell. Getting immediate guidance that would allow parties to avoid litigation sounds like utopia to me.

Third, AI's drawbacks don't matter much in the context of judging.

- AI lacks humans’ creativity—one would not expect AI to write *Infinite Jest* or invent cubism. But for appellate judging, that is fine. “Creative judging” is usually a euphemism for bad judging. If a judge reads the briefs and applies the precedents predictably, the judge is doing a good job.
- Similarly, AI writing is often dull, but that is OK. The primary purposes of judicial opinions are to explain to the parties (who aren’t lawyers) why they won or lost, and for precedential opinions, to explain the law to the world. Clear, dull writing is appropriate for these tasks. I personally enjoy reading cleverly-written judicial opinions, but I am a weird person and that preference is of no social importance.
- In his 2023 year-end report, the Chief Justice wrote: “Judges, for example, measure the sincerity of a defendant’s allocution at sentencing. Nuance matters: Much can turn on a shaking hand, a quivering voice, a change of inflection, a bead of sweat, a moment’s hesitation, a fleeting break in eye contact. And most people still trust humans more than machines to perceive and draw the right inferences from these clues.” Rarely have I heard a better argument for the AI takeover. Years of imprisonment shouldn’t turn on a criminal defendant’s acting skills at sentencing. And abundant theoretical and empirical research suggests that drawing inferences from things like “a fleeting break in eye contact” is an unreliable way of making factual findings. Worse, this type of decisionmaking can be badly infected by cultural bias. If, e.g., a Chadian refugee wearing a hijab breaks eye contact with an immigration judge during an asylum hearing, is the judge really supposed to draw inferences about her credibility from that?

Do the work

This discussion will make AI safety advocates cringe. I realize that if we are concerned about superintelligent AIs turning humans into paper-clips, it is a terrible idea to give AIs control over the legal system. For this reason, I do not advocate turning over our legal system to AI. I have no concern about AI’s ability to decide cases and write judicial opinions in particular cases. But I can think of many science-fiction authors who would have lots of interesting things to say about a society that chooses to place its justice system in the hands of a few tech companies that create AIs whose outputs they do not fully understand.

However, there are many ways to deploy AI in appellate decision-making short of transforming AI into the decision-maker.

The first step is making sure the AI works. We have to know how good AI is at deciding cases, where AI goes wrong, and what prompts should be used to improve AI's performance.

Law professors out there, this shouldn't be that hard to do. Just have your RAs download a few hundred briefs off of PACER, input them into your AI of choice, ask the AI to adjudicate the case, and compare the results to the actual judicial decisions that came down in those cases. If you can convincingly show that AI judges are as good or better than ordinary judges, your article will be cited lots of times. (Suggested title: "The Path of the Law II."). I don't think it would be that hard to create a benchmark that would check whether an AI adjudicator is as accurate as the median human judge, at least in ordinary cases like *Snell* that lack an ideological angle.

Once that is accomplished, there are many options:

- Enforce arbitration awards issued by AI or have law clerks be AI. I've written about this before, and state-of-the-art AI is much better now than then.
- Give litigants the option of having an AI decide their case after their briefs are submitted. Both litigants might prefer to wait 20 seconds rather than 2 years for a final disposition. If we're worried about hallucinations, we can deem those opinions non-precedential.
- Use AI in every case to draft an initial decision, subject to plenary (or maybe abuse-of-discretion) review by the actual judges.
- Use a protocol in which the judges input the bottom-line result and perhaps two lines of reasoning into the AI, and the AI drafts the judicial opinion.
- Use a protocol in which multiple AIs developed by different companies assess the same case; if all of them reach the same outcome, that will be the outcome, whereas if they reach different conclusions, human judges can step in.
- Develop multiple prompts for an individual AI—a textualist prompt, a purposivist prompt, etc.—and assess whether these multiple prompts each produce the same outcome. Again, if varying the prompt leads to varying outcomes, the humans can enter the scene.
- Perhaps we should emphasize AI lawyering more than AI judging? This would solve the problem of misaligned AIs leading humans astray—judges are always skeptical of lawyers' submissions and would likely be even more skeptical of AI lawyers than

human lawyers. Actually, I think AI lawyering is great too, and I dream of the day in which AI briefs are transmitted to an AI adjudicator. They're not mutually exclusive.

There are other options as well. The point is, we should come at this problem in the spirit of "AI has read every case ever written and in most cases, it will be more accurate than humans" rather than "AI is possibly better and possibly worse than the dictionary."