Prediction, Behaviorism and Explanation in Legal NLP: Some Deflationary Thoughts

Dimitrios Tsarapatsanis[γ] **Nikolaos Aletras**[σ]
[γ]University of York [σ]University of Sheffield
dimitrios.tsarapatsanis@york.ac.uk
n.aletras@sheffield.ac.uk

Abstract

Predicting the outcomes of legal cases using natural language processing (NLP) methods has attracted two main objections: (a) that this line of research presupposes a crude *jurisprudence of behaviorism*, which treats adjudication as a black-boxed transformation of inputs into outputs; and (b) that it *conflates prediction and explanation*. In this position paper, we respond to these critiques in a deflationary spirit. First, we argue that predicting judicial outcomes is a mundane and widespread practice which presupposes no behaviorist commitments, and that the claims that can reasonably be made on behalf of NLP systems, including contemporary large language models, are correspondingly modest. Distinguishing *easy* cases from *patterned* and *unpatterned* hard ones, we argue that such systems track the regularities of the first, can help identify the non-legal regularities of the second, and are conjecturally out of reach of the third. Second, we posit that, far from being embarrassed by the distinction between prediction and explanation, computational methods can actually serve as heuristic instruments in the search for proper explanations, in the spirit of the American legal realists. Finally, we provide specific recommendations for the NLP community on how to frame claims of legal predictive systems.

1 Introduction

A decade ago, Aletras et al. (2016) proposed using NLP methods to predict the outcomes of cases decided by the European Court of Human Rights (ECtHR), based solely on textual information extracted from published judgments, a task termed as Legal Judgment Prediction (LJP).

The work inspired numerous subsequent studies in predicting cases of the ECtHR (Chalkidis et al., 2019; Medvedeva et al., 2020; Santosh et al., 2022), the Supreme Court of the United States (Katz et al., 2017), various Courts of China (Xiao et al., 2018; Liu et al., 2023), the Indian Supreme Court (Malik et al., 2021), the Federal Supreme Court of Switzerland (Niklaus et al., 2021), higher Turkish Courts (Mumcuoğlu et al., 2021), the French Supreme Court (Şulea et al., 2017), and UK Courts (Strickson and De La Iglesia, 2020), *inter alia*. While LJP has become a popular NLP task, the community has further proposed more comprehensive benchmarks that go far beyond LJP, including multiple legal tasks such as LexGLUE (Chalkidis et al., 2022), LEXTREME (Niklaus et al., 2023), LegalBench (Guha et al., 2023), LEXAM (Fan et al., 2026), PLAWBENCH (Shi et al., 2026), and ECtHR-NPD (Pu et al., 2026), *inter alia*. Furthermore, specialized Large Language Models (LLMs) in the legal domain have been developed such as Legal-BERT (Chalkidis et al., 2020), LegalLama (Chalkidis et al., 2023), and SaulLM (Colombo et al., 2024).

AI holds the promise to improve access to justice (Zhong et al., 2020), and legal tasks are now included as one of the fundamental domains for evaluating the capabilities of frontier models (e.g., Claude Mythos 5 and Fable 5 developed by Anthropic[1]). However, using AI systems for making predictions about judicial decisions has also attracted criticism on two main grounds: (a) that

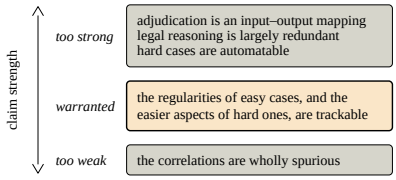


Figure 1: Our position occupies the middle of the LJP claim strength spectrum. Above it lies overreach and below it an underclaim that current systems have falsified.

[1] https://x.com/claudeai/status/2064394151441863006

this line of research presupposes a crude *jurisprudence of behaviorism*, which treats adjudication as a black-boxed transformation of inputs into outputs; and (b) that it *conflates prediction and explanation*.

The first objection, put forward by Pasquale and Cashwell (2018), states that judgment prediction research instantiates what the authors call a *jurisprudence of behaviorism*. The label is borrowed from the behaviorist research program in social psychology associated with B. F. Skinner, which sought to explain human conduct in terms of relations between environmental stimuli and observable responses. Skinner treated the internal workings of the mind of the agent as an explanatorily irrelevant factor. Transposed to judicial behavior, the resulting picture would be one on which judges are treated as *black boxes* converting inputs (e.g., briefs, filings, fact patterns) into outputs (decisions), where prediction of judgments based on the relationship between inputs and outputs is all that matters, to the detriment of reasoned persuasion and justification.

The second critique is that LJP research runs prediction and explanation together. Pasquale and Cashwell (2018) argue that models that are not grounded in the meaning of precedents and facts cannot account for the decisions they anticipate, and that correlations between word frequencies and outcomes may be wholly spurious. Medvedeva et al. (2023) observe that systems described as predicting decisions rarely deliver insights into the reasons behind them. Others have raised empirically grounded concerns. Santosh et al. (2022) show that classifiers trained on ECtHR data are drawn to shallow surface signals which are statistically predictive but legally irrelevant. Chen et al. (2022) put the same point as a choice between the rule of law and what they call *the rule of correlation*, demonstrating that a LJP model could produce correct outputs for the wrong reasons, relying on punctuation marks and function words instead of legal doctrine. Branting et al. (2021) and Bex and Prakken (2021) argue that accuracy is of limited value, unless it is accompanied by explanations tied to legally salient factors. The upshot, for the critics, is that a system which correctly anticipates outcomes tells us little, if anything, about why courts decide cases as they do.

Our reply to both kinds of objections is deflationary. We posit that, properly understood, LJP methods are based on well-respected understandings of the nature of adjudication. What these studies did show is that AI can, but need not, be fruitfully used to track certain past patterns in the case law and to extrapolate from those patterns to hypotheses about future decisions. We contend that this is a real achievement, which bolsters interdisciplinary research that focuses on judicial behavior. However, the results are bounded in both directions. Figure 1 situates our position in the middle of a continuum, beneath overclaims that invite warranted objections, and above an underclaim that we think current systems have already falsified. Our argument is that, once the claims properly made on behalf of LJP are stated with sufficient care, the two objections either dissolve or turn into empirical questions to be settled on their own merits. The contributions of this position paper are as follows:

- We argue that prediction of judicial outcomes is an ordinary, widespread practice presupposing no behaviorist theory of adjudication. Given the familiar jurisprudential distinction between easy and hard cases, the reasonable claim on behalf of AI systems is the modest one that regularities characteristic of easy cases, and possibly (albeit in a more complicated manner) patterned hard cases, can be tracked by computational means (§2);
- We argue that the distinction between prediction and explanation, far from undermining computational work, assigns it a well-defined and potentially illuminating heuristic role in the search for explanations of judicial behavior in patterned hard cases (§3);
- For each of these points, we provide specific recommendations for the community (§4).

2 Prediction Without Behaviorism

2.1 An ordinary professional practice

The charge of a *jurisprudence of behaviorism* assumes that predicting judicial outcomes commits one to the following rather disgraced, theory of adjudication: judges would be reducible to stimulus-response systems, reasoned persuasion would be a mere illusion, and legal reasoning would be the post hoc rationalization of idiosyncratic judicial 'hunches'. But this picture of adjudication does not necessarily follow from judgment prediction projects, either in general or in the particular case of computational prediction.

For one thing, predicting the outcomes of cases is a task that ordinary legal professionals perform

regularly, and more or less successfully, as a routine matter. Consider, for instance, a lawyer advising a client whether to settle, or a legal aid board triaging applications. All of these practitioners engage, in some form or another, in outcome prediction, and none of them thereby adopts, or needs to adopt, a stimulus-response conception of judicial behavior. Indeed, the canonical statement of the predictive stance in Anglo-American jurisprudence, Holmes's idea that law reduces to 'prophecies of what the courts will do in fact' (Holmes, 1897), was offered from the standpoint of the counseling lawyer. One can predict outcomes while also believing that judges are conscientious reasoners responding predominantly to well-defined and determinate legal reasons governing particular cases. That belief, incidentally, actually explains why prediction succeeds so often and, more generally, how the legal system produces certainty and thereby protects people's expectations about how the law will be applied to their conduct. Judicial reasoning under shared, determinate, and public standards is a powerful generator of regularity.

Hence, *prediction is a widespread practice which isolates certain factors deemed pertinent, and there is no valid inference from 'X predicts judicial outcomes' to 'X treats judges as black boxes'*.

2.2 Prediction and reasons

The point can be sharpened further. Hart (1961) famously objected that predictive theories of law only capture the external stance of the observer, who records the behavioral regularities of officials, while missing out the internal point of view of participants, who treat legal standards as reasons for acting and as grounds for criticizing others in cases of non-compliance. Hart observed that a judge who wants to find out what the law is does not predict her own future decision but instead uses legal reasons to reach it.

Still, whilst Hart's argument foregrounds the problems of defining law as a prediction exercise, it is not an objection to predicting outcomes per se. In fact, the adoption of a Hartian internal point of view to the legal system can help predict outcomes in easy cases. In Poscher (2026) terms, these would be cases that require straightforward rule-following. But, more fundamentally, there are also hard cases, mainly appellate ones, in which legal reasons 'run out', the law appears indeterminate, legal experts, including judges, disagree (Poscher, 2026) and successful prediction, when it is possible at all, rests on identifying operative, and patterned, non-legal reasons such as, for example, judges' political attitudes or sense of justice (Ruger et al., 2004). This is, precisely, the explanatory program that was first articulated by traditional American legal realism and has more recently been reformulated and defended by Leiter (2007). The program is not behaviorist in the strict sense, since it explains outcomes through reasons and does not assume that judges are 'black boxes' responding to stimuli. But it goes beyond Hart insofar as it explores how non-legal reasons of the kind alluded to above shape these outcomes beyond the requirements of legal doctrine.

It bears noting that the predictive research program spawned by American legal realism has been enormously influential. Consider, for example, the rich empirical literature on judicial decision-making, including the study of panel effects, of the influence of oral argument, and of the determinants of certiorari, none of which is ordinarily thought to rest on an austere Skinnerian psychology (Heise, 2002). The same, of course, also applies to the routine advisory work of practicing lawyers to which we have already alluded. Since nobody supposes that these activities commit their practitioners to an austere black-box theory of the judicial mind, and indeed practitioners themselves predict outcomes by attempting to identify, in each case, operative sets of legal and/or non-legal reasons, the objection to the effect that prediction equates behaviorism overshoots accordingly.

In short, *predictive approaches are fully compatible with reason-based explanations of judicial behavior, whether the reasons are legal and identified from a Hartian internal point of view, or non-legal and identified from a perspective external to the legal system*.

2.3 Easy cases, hard cases and the deflated claim

What, then, is the reasonable claim to make on behalf of computational outcome prediction? We begin answering this question by first observing that the claims made in LJP work are modest, typically hedged by several caveats and the field has progressed incrementally by taking on board various pertinent critiques. For example, we have already acknowledged that the critics are right that spurious textual correlations lack explanatory value. So a basic desideratum of high-quality research in the field is to find ways of smoking such correlations

out. Precisely in this vein, Chen et al. (2022) have provided a set of techniques and tools that allow researchers to key predictive models to genuine legal reasons driving judicial decision-making easy cases, rather than spurious statistical correlations. Similarly, the construction of interdisciplinary research teams comprising legal experts along with computer scientists can go a long way towards parrying the spurious correlation risk. Relatedly, Chalkidis et al. (2021) extract paragraph-level rationales for ECtHR outcome predictions, so that the textual basis of a prediction can be inspected and tested against the reasoning of the case.

This kind of research design was already in place in Aletras et al. (2016), and in many subsequent instances of LJP literature. There, the features generated by the predictive models were assessed by legal experts against legal criteria, and many of them, though not all, were interpreted as capturing not just statistical artifacts, such as spurious correlations between words, but patterns of easy cases and the determinate, shared legal reasons that explain them. When the correlations could not be rationalized by reference to established legal doctrine, this was openly recognized and thematized as an important limitation of the approach. But it is equally crucial to emphasize that even these early approaches clearly showed that easy, repetitive, cases could be computationally predicted.

We have already mentioned the crucial and familiar jurisprudential distinction between easy and hard cases (Hart, 1961; Dworkin, 1977; Poscher, 2026) in the previous subsection, and we think that more systematic reflection on it can go a long way towards improving the conceptual parameters of the debate on LJP. Recall that, in easy cases, the applicable standards, the facts and their subsumption relations are settled and determinate enough to allow legal experts to converge on a single legally warranted outcome. Hence, these cases exhibit patterns explained by reference to legal reasons. In hard cases, by contrast, the settled legal materials, such as legislative texts and past decisions, are perceived by legal experts to underdetermine the outcome in the sense that both a certain legal solution and a different one, perhaps even its opposite, can be equally well legally justified (Leiter, 2007). This in itself, however, does not mean that there are no explanatory patterns. As the American legal realists fist showed, these just depend, when they can be genuinely identified, on the non-legal reasons that guide judges to pick a certain solution to a given dispute as opposed to another, and to which we have already alluded. Indeed, the whole research program inaugurated by the American legal realists consisted, precisely, in identifying patterns in judgments systematically deviating from 'established legal doctrine' in order to ask which kinds of non-legal reasons explained them (Leiter, 2007).

To discharge the explanatory task, the realists turned their attention to regularities in judicial behavior, tied to fact-patterns, situation-types and institutional context. The job of a genuinely empirical science of law was, precisely, to uncover these regularities and to smoke out the non-legal reasons to which the judges responded (Llewellyn, 1930). So, for example, American realists found that cases by appellate courts in disputes involving employers and employees could be explained not by the standard doctrinal categories of contract law, including breach of contract, but, rather, by the judges' 'sense of justice', which prompted them to use the doctrinal indeterminacy of contract law (i.e., the fact that a legitimate legal argument could be made both in favor of one kind of outcome and for its opposite) to find disproportionately in favor of employees.

This, however, does not imply that such patterns could always be found in hard cases. Sometimes, courts partly engage in what can be described as a creative exercise in forward-looking conceptual engineering, specifying, and modifying the legal concepts in play or even, in rare but well-known and dramatic cases, ushering in conceptual innovations. Seminal well-known examples from US constitutional law include the *Brown v. Board of Education* 1954 Supreme Court case, in which the Court reversed its past case law to decide that racial segregation in schools was in violation of the 14th Amendment of the US Federal Constitution. But even apart from such extraordinary examples, hard cases may be decided by simply taking one horn of a given legal dilemma as a matter of either commitment or drift, and where the decision is not explained by legal or non-legal reasons but is simply a response to a hard choice (Chang, 2024). In view of this, we distinguish between easy cases (where the patterns are explained by determinate, shared, legal reasons), patterned hard cases (where the patterns are explained not by underdeterminate legal reasons but by determinate and shared non-legal reasons that explain judicial responses) and unpatterned hard cases (where conceptual engineering, conceptual innovation or some other form of com-

mitment or drift under conditions of hard choice call the shots).

Now, suppose that a version of this distinction between different types of cases is granted. The claim that AI systems achieve non-trivial accuracy, especially where such systems capture patterns of determinate legal reasons in their features, can then deflate, at least as regards easy cases, into the claim that the regularities characteristic of such cases can be faithfully tracked by statistical models of text. As we have already said, that was, in substance, one of the main claims typically made in earlier LJP studies (Aletras et al., 2016; Xiao et al., 2018) that prompted the discussion on computational prediction of judicial outcomes. We think that this claim is now even better supported by the evidence that has accumulated in the meantime. For one thing, the pattern of benchmark results tracked by later studies is precisely the one predicted by the distinction between easy and hard cases. Systems perform comparatively well on easy cases with determinate correct legal answers and are challenged by open questions demanding structured, multi-step legal reasoning with potentially indeterminate answers (Fan et al., 2026; Shi et al., 2026). For another, methods now exist for tying a prediction to a specific textual basis that would be recognizable by a legal expert, i.e., a paragraph-level rationale (Chalkidis et al., 2021; Chen et al., 2022).

Of course, these results do not automatically dispel the spurious correlation risk. For example, Valvoda and Cotterell (2024) find that models predict ECtHR outcomes well while relying on precedent quite differently from human judges, and the artifact findings noted above (Santosh et al., 2022) should mitigate any unearned enthusiasm for automatic LJP. But our claim is narrower. It consists in saying that, with the right experimental design, the regularities of easy cases are trackable.

Things, however, are different when it comes to hard cases. In fact, with respect to at least unpatterned such cases, one could perhaps mount a principled argument for cabining more ambitious predictive claims involving the use of LLMs. LLMs are, by construction, sophisticated statistical continuations of past textual patterns, in the sense that their outputs are high-probability continuations of inputs given a training distribution (regardless of whether this is either pre- or post-training data). While current generative models can generalize to unseen cases that have similarities to cases seen during training, unpatterned hard cases often prompt conceptual change and, occasionally, genuine conceptual innovation on the part of judges. Alternatively, hard cases can be those where hard choice under parity of opposite legal arguments simply makes any choice rationally arbitrary. Rather than repeating past patterns, judges in these circumstances ask questions about how concepts should be modified and attempt to tailor these concepts to new unmet needs within ongoing social practices (Queloz, 2025) or commit to certain features of the cases as the significant ones in a rationally underdetermined way (Chang, 2024).

There is reason to doubt that such innovation or hard choice can be performed by existing LLMs operating alone. The empirical picture, so far as it goes, seems to corroborate important, and perhaps insuperable, difficulties in this area. For example, on LEXAM, models are challenged specifically by open questions requiring structured, multi-step legal reasoning (Fan et al., 2026), and on PLAW-BENCH none of the systems evaluated performs strongly on tasks drawn from real legal practice (Shi et al., 2026). Outside law, heavy reliance on generative systems has been found to flatten the space of novel content (Doshi and Hauser, 2024), though that evidence concerns creative writing. If this is right, then an aggregate accuracy figure, insofar as it incorporates easy/determinate cases, tells us nothing about the automation of adjudication in its creative or hard moments. A single number over a docket mixing different kinds of case cannot report performance on the hard cases in which human judgement matters most. More importantly, for our purposes, we think that high accuracy alone could simply not be leveraged to this effect. We shall return, moreover, to the more contested and tractable issue of patterned hard cases in the next section.

On the other hand, we do not want to be understood as establishing by conceptual fiat some kind of inherent ceiling. Benchmark difficulty at any given moment is not evidence of an insuperable limit, and an argument from training objectives does not settle what a trained system can do. We therefore advance it as a conjecture with a falsification condition. Our hypothesis will be falsified if systems come to perform well on unpatterned hard cases under evaluations based on human legal experts' annotations that score well on both the reasoning process and the outcome, and that draw on disputes postdating the training data.

Two upshots follow. First, as already noted, the

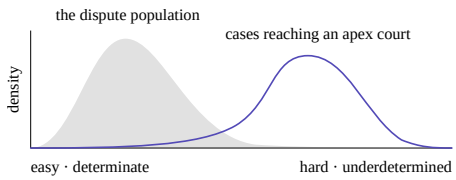


Figure 2: Where selection operates strongly, it displaces the cases reaching an apex court towards the underdetermined end, on which LJP accuracy is typically measured. The magnitude may vary by court. Curves are schematic and normalized independently.

behaviorism charge is too wide. A deflated predictive claim about tracked regularities in easy cases is compatible with virtually any respectable theory of adjudication, including robustly rationalist ones. Insofar as NLP systems track easy cases and extrapolate successfully from them, they also tell against the stronger version of the objection at the other extreme, on which the correlations between textual features and outcomes are wholly spurious (Pasquale and Cashwell, 2018).

Second, the deflated claim should at the same time give pause to the more enthusiastic proponents of potential predictive uses of LLMs. Reported accuracy figures are estimates of pattern-traceability over particular case distributions. Researchers should pay lots of attention, aided by legal experts, on the types of cases that data sets comprise. For example, distributions can be shaped by selection effects of the kind identified by Priest and Klein (1984), whereby the cases reaching a given court can be systematically unrepresentative and disproportionately underdeterminate. Alternatively, the dataset might comprise many easy/repetitive cases. Figure 2 illustrates the resulting displacement where the distribution on which accuracy sits further towards the underdetermined end of the spectrum than the population of disputes from which it is drawn. The distributive structure of any given docket is an empirical matter, and it varies with the court. For instance, where review is discretionary, selection operates strongly and hard/indeterminate cases would tend to be over-represented in the docket. Alternatively, where a court must dispose of a high volume of repetitive applications, easy cases would tend to predominate. Thus, neither profile can be simply assumed from a court's place in the hierarchy, which is why we recommend in §4 that accuracy claims be indexed to the distribution over which they are actually measured. To repeat, whereas patterned hard cases can be, under conditions that we explore in §3.3, fruitfully investigated using LLMs, we strongly doubt that the creative legal reasoning manifested in unpatterned hard cases is LLM-replicable.

2.4 What the LJP critics get right

None of this is to say that the concerns animating the behaviorism charge are misplaced. Charitably understood, they are legitimate concerns not about prediction per se, but about rule of law values. Pasquale and Cashwell (2018) are principally worried about the potential risks of deployment of legal-predictive NLP, to wit, about risks that could follow if predictive systems were actually used by officials, including judges, to sort, triage or decide cases, displacing the giving of reasons with statistical sorting that those affected cannot realistically contest. That worry is serious and we share it to the fullest possible extent. It is, however, a worry about the institutional uses to which the technology is put, not about the legitimacy of a descriptive research program, and the two should not be confused. The same concern has been pressed from within AI and Law. Bex and Prakken (2021) and Branting et al. (2021) argue that a sole outcome prediction, however accurate, gives a judge nothing on how the case should be decided, and that such a system could be useful in a case only if it supplied a justification for its prediction.

The regulatory (Binns, 2022; Goanta et al., 2023) and ethical (Leins et al., 2020; Tsarapatsanis and Aletras, 2021; Valvoda et al., 2024) aspects of NLP applied to legal text and more broadly deserve separate and sustained treatment. Running together the two questions, however, hinders lucid debate. *On the one hand, it leads critics to attack research designs when their real target is institutional design. On the other hand, it prompts proponents to answer institutional objections with accuracy figures, which simply beg the interesting questions. Both of these tendencies should be resisted.*

3 Prediction and Explanation

3.1 A commonsense distinction

The second objection insists that prediction is not explanation. We agree, and we regard the point as common sense, even prosaic. We do not view it as a significant discovery. A model that anticipates outcomes from textual features does not thereby explain why courts decide cases the way they do. In other words, correlational success underdetermines

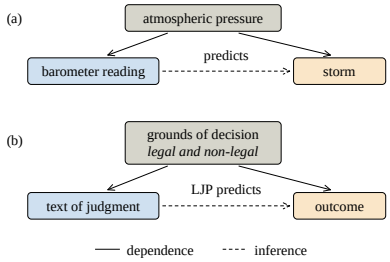


Figure 3: Top (a): atmospheric pressure produces both the barometer reading and the storm, so the reading predicts the storm along a path that carries no explanatory weight. Bottom (b): the model infers the outcome from the text, whereas the dependence that would explain the outcome runs from the grounds of decision, which the model does not represent. The asymmetry is a general feature of empirical inquiry not a shortcoming of LJP.

causal and rational structure, and post-hoc inspection of predictive features is a notoriously fallible guide to underlying causal mechanisms (Lipton, 2018; Jacovi and Goldberg, 2020; Santosh et al., 2022).

3.2 The asymmetry between prediction and explanation

The relation between prediction and explanation has been examined at length outside law, and the conclusion reached there is precisely the one the critics urge. Hempel and Oppenheim (1948) proposed that explanation and prediction share a single logical structure, so that any adequate explanation could have served as a prediction and vice versa. This thesis did not survive subsequent scrutiny. Scriven (1962) offered the standard counterexample, to the effect that a rapidly falling barometer reliably predicts a storm but plainly does not explain it. Bromberger (1966) added the case of the flagpole, whose height can be derived from the length of its shadow without the shadow explaining the height. The contemporary widely shared diagnosis is that explanation requires information about dependence, typically causal dependence theoretically articulated, which mere predictive correlation does not supply (Woodward, 2003; Douglas, 2009).

Note, moreover, that, insofar as the asymmetry between prediction and explanation is a general feature of empirical inquiry, it is not a specific objection to legal NLP. Barometers are not discredited by the observation that they do not explain storms. It is simply the case that the information that may be inferred from predictive success is not, per se, adequate for explanation absent a specification of the causal mechanisms that produce the outcome (Figure 3). Moreover, we have already stated, when it comes to judicial behavior, these mechanisms are not always reducible to the operation of determinate legal reasons specifiable ex ante, since, in hard cases, strictly legal reasons inferred from conventional legal sources underdetermine outcomes.

3.3 Computational methods in the search for explanations

Still, the distinction between types of cases, properly drawn, assigns predictive modeling a positive role. In empirical inquiry generally, robust predictive regularities are not themselves explanations, but they can prepare the ground for seeking them, insofar as they indicate where stable structure lies and thus where explanatory effort should be directed. That was, it bears repeating, the aspiration of the American legal realists, who placed particular emphasis on stable decisional patterns (Leiter, 2007) (§2.3).

We can now return to the category of *patterned* hard cases, i.e., cases where judges reach outcomes on the basis of patterns of non-legal reasons, since legal reasons are indeterminate since they point to equally legitimate but potentially conflicting directions. Our argument is that NLP methods and LLMs have a distinctive contribution to make, precisely because they can point to explanatory factors in large corpora of legal text that may resist articulation in accepted legal-doctrinal terms.

Where a model achieves predictive success that cannot be reconstructed from the doctrinally recognized and widely shared legal reasons (which means that this is not an easy case), the success itself is an indication that there may exist an appropriate explanatory target. The target consists in regularities, whether in facts, framing, procedural posture or institutional practice, that the official self-description of the practice renders invisible.

The subsequent explanatory work, formulating and testing hypotheses about the causal mechanisms that generate the regularities, remains a task for lawyers and legal theorists. Whilst it may be that no convincing mechanistic explanation can be provided (in cases of conceptual innovation or simply because the mechanisms cannot be unearthed), at the very least predictive success can be a useful data point for identifying perhaps surprising patterns, just like the American legal realists did.

It bears repeating, however, that this does not entail that *any* regularity will do. Regularities are

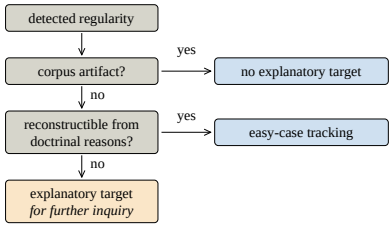


Figure 4: Predictive success identifies an explanatory target only where the regularity is genuine and resists reconstruction from doctrinal reasons.

only worth further exploring if they are genuine, and the legal NLP literature gives ample reason for caution on that score. As we have already insisted, where the signal a model exploits turns out to be an artifact of corpus construction or of surface features of the text (Santosh et al., 2022; Chen et al., 2022), there is no explanatory target at all. The methodological moral is that the heuristic use of LJP models presupposes some prior confidence that what has been detected is a property of the relevant judicial practice without being an attribute of the dataset (Figure 4). Such confidence is easier to secure with interpretable models than with post-hoc rationalizations (Rudin, 2019).

Nevertheless, an objection to the points we make above is that this division of labor has been overtaken by events. Contemporary systems are prompted or trained to reason step by step (Wei et al., 2022). They produce deductive rationales that cite provisions, distinguish precedents and weigh factors, and so appear to supply just the doctrinal reconstruction we have said predictive models cannot give. The objection does not survive the asymmetry of §3.2. A reasoning trace is a further output of the same system, not evidence about the dependence structure that produced the outcome, and whether it describes that structure is an empirical question that has been asked and answered discouragingly.

Models produce plausible rationales that omit the features actually driving their predictions (Turpin et al., 2023; Lanham et al., 2023), and reasoning models verbalize decisive influences in a minority of the cases where those influences demonstrably operate (Chen et al., 2025). In law specifically, fluent legal prose is a poor guide to whether the authorities it invokes exist (Dahl et al., 2024). A doctrinally articulate trace is precisely what would lead an investigator to answer *yes* at the second node of Figure 4 and conclude that no explanatory target remains, and it can do so whether or not the underlying computation was sensitive to the doctrine it recites. Reasoning traces are thus a new way of failing it the triage, and should be better treated as hypotheses about the model, to be tested by intervention (Jacovi and Goldberg, 2020).

4 Recommendations

Our arguments in §2 and §3 converge on a single methodological directive: *these debates should be conducted at the level of specific texts, specific tasks and specific claims, where they can actually be adjudicated*. Table 1 pairs framings we take to overreach with the corresponding claims the evidence supports. Our recommendations are as follows:

Index claims to the easy/hard distinction. Claims made on behalf of LJP systems should be indexed explicitly to the distinction between easy and hard cases, and to the selection effects shaping the distributions over which accuracy is measured (Figure 2). A reported accuracy of a model is an estimate of pattern-traceability over a particular docket, not a measure of the redundancy of legal reasoning. Legal experts should always be ideally involved to help distinguish between easy and hard cases.

Do not inflate the deflated claim. Both critics and proponents should resist converting a modest claim about the computational traceability of adjudicative regularities into a substantive behaviorist theory of adjudication, whether for purposes of endorsement or of condemnation. The deflated claim is compatible with virtually any respectable theory of adjudication, including robustly rationalist ones.

State deployment objections as such. Objections on the use of predictive systems by officials, including judges, are objections to institutional design based on rule of values. We posit that these are not directly relevant to a descriptive research program. They are serious, and they deserve separate treatment and analysis on their own terms beyond quantitative metrics.

Predictive results as explananda. LJP models should be seen as instruments in the search for explanations of judicial behavior, not as explanations in themselves (Figure 3). This holds for rationales such systems now generate. For example, a reasoning trace is a further output of the model without necessarily being evidence about what produced

Inflated framing	Deflated framing
Claims made on behalf of LJP research	
The system predicts judicial decisions.	The system tracks outcome-correlated regularities in the case materials it is given.
Macro-F_1 of 0.79.	Macro-F_1 of 0.79 over the case distribution this corpus samples, which selection effects have shaped, not a measure of the redundancy of legal reasoning.
Strong benchmark scores show that the model can perform legal reasoning.	Strong benchmark scores show that patterns well represented in the training distribution can be accurately continued.
The features, or the model's stated reasoning, reveal the determinants of decision.	Either indicates candidate explananda, whether it describes what produced the outcome is a further empirical question.
Readings imposed by LJP critics	
Judgment prediction treats adjudication as a black-boxed mapping from inputs to outputs.	Judgment prediction is compatible with reason-explanations of judicial behavior, legal and non-legal.
Predictive systems displace the giving of reasons.	An objection to institutional deployment, which we share, and not to the legitimacy of a descriptive research program.

Table 1: Framings by LJP proponents and critics that overreach and the corresponding claims the evidence supports.

the outcome. Reporting it with a prediction does not discharge the explanatory task. Where predictive success cannot be reconstructed from doctrinally recognized reasons, the model has located an explanatory target. The evidence for that judgment should be reported, since the system's own account of its reasoning does not by itself supply it. Formulating and testing hypotheses about the mechanisms that generate the regularity remains a task for legal experts and socio-legal scholars.

Separate the practice from the corpus. Before a regularity is treated as an explanatory target, it should be distinguished from artifacts of corpus construction and of surface features of the text (Figure 4). Such confidence is easier to secure with interpretable models than with post-hoc rationalizations of opaque ones (Rudin, 2019).

5 Conclusion

We have replied to two objections to the LJP research. Our strategy has been deflationary throughout. The charge of behaviorism fails because prediction is an ordinary professional practice carrying no theoretical commitments about adjudication. Adopting the external point of view for particular purposes does not deny the internal point of view of participants in legal practices and is fully compatible with reason-based explanations of judicial behavior. The reasonable computational claim that the regularities of easy cases can be tracked is modest. It extends, more tentatively, to patterned hard cases, where computational methods can point to non-legal regularities that doctrinal self-description does not register. It does not extend to unpatterned hard cases, for which we have offered reasons, as a conjecture, to doubt that prediction lies within reach of LLMs. The distinction between prediction and explanation is well established and applies across empirical inquiry. Properly drawn, it assigns models the role of detecting patterns, invisible to doctrinal self-description, which can prepare the ground for explanation.

Overall, our recommendations converge on a single methodological directive: *that these debates be conducted at the level of specific texts, specific tasks and specific claims, where they can actually be adjudicated*.

Limitations

First, we take the easy/hard distinction as a working jurisprudential category. The subsequent distinction between patterned and unpatterned hard cases is our own. Operationalizing either, whether by expert annotation or by proxies such as dissent, panel disagreement, court level or reversal, is a substantive empirical task (beyond the scope of this paper) that our first recommendation makes pressing. The second distinction is the harder case. Whether a hard case is patterned is not settled in advance, but is established, if at all, by the inquiry we describe in §3.3. The taxonomy is better read as a research agenda than as a classification scheme that is ready for use.

Second, the claim that conceptual innovation lies beyond LLMs is advanced in §2.3 as a conjecture. The benchmark evidence we cite establishes present difficulty and the faithfulness results we rely on in §3.3 are drawn largely from outside law. Third, our illustrative examples come predominantly from common-law jurisdictions and the ECtHR, and both the easy/hard distinction and the realist framing of §2.3 have Anglo-American origins. How applicable they are on other legal

traditions is an open avenue for future research.

Finally, we set aside the regulatory and ethical questions raised by deployment, for the reasons given in §2.4. Our recommendation there concerns how such objections should be classified, not how they should be answered. Answering them requires separate treatment.

Acknowledgments

NA is partly supported by the EPSRC [grant number EP/Y009800/1], through funding from Responsible AI UK (KP0016) as a Keystone project.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93.

Floris Bex and Henry Prakken. 2021. On the relevance of algorithmic decision predictors for judicial decision making. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 175–179.

Reuben Binns. 2022. Human Judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & governance*, 16(1):197–211.

L. Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2021. Scalable and explainable legal prediction. *Artificial Intelligence and Law*, 29(2):213–238.

Sylvain Bromberger. 1966. Why-questions. In Robert G. Colodny, editor, *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, pages 86–111. University of Pittsburgh Press.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Ruth Chang. 2024. Human in the Loop! In David Edmonds, editor, *AI Morality*, chapter 20, pages 222–234. Oxford University Press, Oxford.

Haotian Chen, Lingwei Zhang, Fanchao Chen, and Yang Yu. 2022. Knowledge is power: understanding causality makes legal judgment prediction models more generalizable and robust. *Preprint*, arXiv:2211.03046.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. 2025. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*.

Pierre Colombo, Telmo Pires, Malik Boudiaf, Rui Melo, Dominic Culver, Etienne Malaboeuf, Gabriel Hautreux, Johanne Charpentier, and Michael Desa. 2024. Saullm-54b & saullm-141b: Scaling up domain adaptation for the legal domain. *Advances in Neural Information Processing Systems*, 37:129672–129695.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Anil R. Doshi and Oliver P. Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290.

Heather E. Douglas. 2009. Reintroducing prediction to explanation. *Philosophy of Science*, 76(4):444–463.

Ronald Dworkin. 1977. *Taking Rights Seriously*. Harvard University Press.

Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, et al. 2026. Lexam: Benchmarking legal reasoning on 340 law exams. In *International Conference on Learning Representations*, volume 2026, pages 43451–43489.

Catalina Goanta, Nikolaos Aletras, Ilias Chalkidis, Sofia Ranchordás, and Gerasimos Spanakis. 2023. Regulation and NLP (RegNLP): Taming large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8712–8724, Singapore. Association for Computational Linguistics.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, K. Aditya, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Herbert L. Adolphus Hart. 1961. *The Concept of Law*. Clarendon Press.

Michael Heise. 2002. The Past, Present, and Future of Empirical Legal Scholarship: Judicial Decision Making and the New Empiricism. *University of Illinois Law Review*, 2002(4):819–850.

Carl G. Hempel and Paul Oppenheim. 1948. Studies in the logic of explanation. *Philosophy of Science*, 15(2):135–175.

Oliver Wendell Holmes. 1897. The path of the law. *Harvard Law Review*, 10(8):457–478.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205.

Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, 12(4):e0174698.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Brian Leiter. 2007. *Naturalizing Jurisprudence: Essays on American Legal Realism and Naturalism in Legal Philosophy*. Oxford University Press.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43.

Yifei Liu, Yiquan Wu, Yating Zhang, Changlong Sun, Weiming Lu, Fei Wu, and Kun Kuang. 2023. Mlljp: Multi-law aware legal judgment prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1023–1034, New York, NY, USA. Association for Computing Machinery.

Karl Llewellyn. 1930. A realistic jurisprudence — the next step. *Columbia Law Review*, 30(4):431–465.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2020. Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, 28:237–266.

Masha Medvedeva, Martijn Wieling, and Michel Vols. 2023. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31:195–212.

Emre Mumcuoğlu, Ceyhun E Öztürk, Haldun M Ozaktas, and Aykut Koç. 2021. Natural language processing in law: Prediction of outcomes in the higher courts of turkey. *Information Processing & Management*, 58(5):102684.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A multi-lingual and multi-task benchmark for the legal domain. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3016–3054, Singapore. Association for Computational Linguistics.

Frank Pasquale and Glyn Cashwell. 2018. Prediction, persuasion, and the jurisprudence of behaviourism. *University of Toronto Law Journal*, 68(supplement 1):63–81.

Ralf Poscher. 2026. The autonomy of legal doctrine: its merit in the process of social differentiation. In *Oxford Studies in Philosophy of Law*, pages 63–98. Oxford University Press.

George L. Priest and Benjamin Klein. 1984. The selection of disputes for litigation. *The Journal of Legal Studies*, 13(1):1–55.

Yanyi Pu, Damian Gonzalez-Salzberg, Zheng Yuan, and Nikolaos Aletras. 2026. How much is a human right worth? ECtHR-NPD: A benchmark for predicting non-pecuniary damage awards. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 30138–30165, Budapest, Hungary. Association for Computational Linguistics.

Matthieu Queloz. 2025. *The Ethics of Conceptualization: Tailoring Thought and Language to Need*. Oxford University Press.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Theodore W. Ruger, Pauline T. Kim, Andrew D. Martin, and Kevin M. Quinn. 2004. The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decision-Making. *Columbia Law Review*, 104(4):1150–1209.

T. Y. S. S. Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022. Deconfounding legal judgment prediction for European Court of Human Rights cases towards better alignment with experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1138.

Michael Scriven. 1962. Explanations, predictions, and laws. In Herbert Feigl and Grover Maxwell, editors, *Scientific Explanation, Space, and Time*, volume 3 of *Minnesota Studies in the Philosophy of Science*, pages 170–230. University of Minnesota Press.

Yuzhen Shi, Huanghai Liu, Yiran HU, Song Gaojie, Xu Xinran, Yubo Ma, Tianyi Tang, Li Zhang, Qingjing Chen, Feng Di, Wenbo Lv, Weiheng Wu, Kexin Yang, Sen Yang, Wei Wang, Rongyao Shi, Qiu Yuanyang, Yuemeng Qi, Zhang Jingwen, and 11 others. 2026. PLAWBENCH: A rubric-based benchmark for evaluating LLMs in real-world legal practice. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10067–10116, San Diego, California, United States. Association for Computational Linguistics.

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal judgement prediction for uk courts. In *Proceedings of the 3rd International Conference on Information Science and Systems*, pages 204–209.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the law area and decisions of French Supreme Court cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria. INCOMA Ltd.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Josef Valvoda and Ryan Cotterell. 2024. Towards Explainability in Legal Outcome Prediction Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7269–7289, Mexico City, Mexico. Association for Computational Linguistics.

Josef Valvoda, Alec Thompson, Ryan Cotterell, and Simone Teufel. 2024. The Ethics of Automating Legal Actors. *Transactions of the Association for Computational Linguistics*, 12:700–720.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

James Woodward. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.