The Ethics of Automating Legal Actors

Josef Valvoda[Φ]   **Alec Thompson**[Φ]   **Ryan Cotterell**[Ψ]   **Simone Teufel**[Φ]

[Φ]University of Cambridge, UK   [Ψ]ETH Zürich, Switzerland

{jv406,at808,sht25}@cam.ac.uk   ryan.cotterell@inf.ethz.ch

Abstract

The introduction of large public legal datasets has brought about a renaissance in legal NLP. Many of these datasets are composed of legal judgments—the product of judges deciding cases. Since ML algorithms learn to model the data they are trained on, several legal NLP models are models of judges. While some have argued for the automation of judges, in this position piece, we argue that automating the role of the judge raises difficult ethical challenges, in particular for common law legal systems. Our argument follows from the social role of the judge in actively shaping the law, rather than merely applying it. Since current NLP models are too far away from having the facilities necessary for this task, they should not be used to automate judges. Furthermore, even in the case that the models could achieve human-level capabilities, there would still be remaining ethical concerns inherent in the automation of the legal process.

1 Introduction

This paper discusses the ethical aspects of using natural language processing (NLP) research to augment or replace the work of legal experts, with particular emphasis on common law legal systems. Although we agree that there are potential benefits to the practical application of NLP to the legal domain, these applications face several ethical challenges. Some of these challenges are resolvable with technical advances; others, however, appear to be intrinsic to using any kind of automation. We consider the two main legal actors, the judge and the lawyer, and find that while automation of either can be beneficial, lawyer automation presents fewer challenges. However, current work in legal NLP and the legal domain are motivated by modeling judges directly: Many current proposals for automation in the legal domain concentrate on the judiciary, and similarly, a number of popular legal datasets represent text produced by judges.

We begin by giving a brief legal background in §2, where we explore the role of judges and lawyers in common law legal systems and their distinct functions. We then introduce the ideas of the rule of law and substantive justice—the pillars on which the judicial system is built and which connect law and morality. In §3, we explore practical automation proposals discussed in legal NLP research. These proposals can be broken into two groups: The first of these advocate a complete replacement of legal professionals with technology, whereas the second group advocates a mere augmentation of judges or lawyers by partial automation or by supplementation of legal tasks with NLP. Concluding our background sections, in §4, we turn to legal NLP and discuss the impact of the shift from symbolic to sub-symbolic AI on the field.

We then proceed in three stages. In the first stage, §5, we outline the risks and benefits of legal NLP. We suggest that current proposals for implementing legal NLP face three technical challenges, namely, (1) the lack of contextual and social intuition at the trial stage; (2) the inability to make controversial moral and political decisions to develop the law; and, (3) the inability to justify the decisions to the public. Further, even if these technical challenges are resolved, existing proposals pose unavoidable ethical risks. They have the potential to (1) centralize power; (2) produce a more brittle legal system; and (3) undermine the accountability of policymakers.

In the second stage (§6), we turn to the main legal actors, the judge and the lawyer, to evaluate their exposure to the risks and benefits. We show that the various concerns affect lawyers and judges differently, and we come to the conclusion that the more promising domain for legal NLP is the role of the lawyer. In a nutshell, lawyers are not burdened with the task of being lawmakers. Instead, the success of their work can be measured more easily, namely, whether or not they can convince judges and other lawyers of the quality of their arguments.

In the third stage (§7), we turn to the role of the judge in contemporary legal NLP research. Many existing legal NLP datasets comprise collections of cases, in other words, text produced by judges. We demonstrate how a lack of data about lawyers and their arguments can undermine the modeling task based on these datasets. To mitigate this issue, we recommend that legal NLP researchers should focus more on the role of the lawyer in the legal system. In the conclusion (§8), we contextualize our work in the wider NLP ethics discussion.

Our proposals do not call for a wholesale re-organization of the field. If we want to further our understanding of law and language, judges remain an important subject to study. Some datasets already implicitly contain the voice of a lawyer, in the form of legal claims, although these need to be disentangled from the voice of the judge. For example, Chalkidis et al. (2021) and Valvoda et al. (2023) predict both lawyers' claims and judicial outcomes, using the same legal dataset, an early step in the right direction, and a demonstration that the shift to the lawyer's perspective can sometimes be achieved by a simple re-purposing of existing datasets. We therefore believe that focusing on the lawyer and their interactions with the judge can supplement existing approaches, and inspire new, more robust legal NLP research.

2 Legal Background

Our work primarily focuses on the common law system practiced in countries such as England, the US, Australia, and India. Civil law, the other major legal system, is practiced in countries such as France, Germany, Japan, and China (Zweigert and Kötz, 1992). Civil law countries rely predominantly on the statutes for legal interpretation. Common law countries, on the other hand, create legal norms through past court decisions (i.e., precedent). Note that both systems utilize statutory and precedential reasoning. However, only in common law does the precedent become binding law.

We focus on the common law system for two reasons. First, common law has a longstanding philosophical debate surrounding it. Second, practical proposals for NLP have mostly focused on common law (Cobbe, 2020), and many of the datasets discussed in §4 and §7 involve common law material. However, we acknowledge there is

much work in legal NLP applied to civil law countries. Chinese law, in particular, has been studied extensively through the lens of NLP tools over the years (Wang et al., 2019; Shen et al., 2020; Yao et al., 2022).

2.1 Legal Actors

Under both systems, there is a strict distinction between the role of judges and lawyers. Members of the public go to lawyers to frame their needs in legal terms, which get expressed in the form of legal documents, such as wills, deeds, and contracts. A lawyer might also be asked to assist if the client wants to sue someone or is being sued. The client specifies their desired outcome to the lawyer, such as avoiding responsibility for causing an injury, and lawyers then translate these demands into legal claims. A claim is a legal argument which suggests that the client's preferred dispute outcome is consistent with the law (Sako et al., 2022). These claims are usually sent to the opposing party's lawyers in the form of a legal brief before there is any litigation. The most common result is for the parties to settle at this point, preventing the case from going to trial (Sturge, 2021). If the parties decide to take the case to court, the arguments will be repeated in front of a judge.

The central duty of the lawyer is to their client. They are responsible to the public only to the extent they commit no crimes and fulfill their professional obligations.[1] Judges, on the other hand, are not hired by members of the public. Instead, they work for the state and have a duty to the public to decide cases correctly and fairly (Wacks, 2015). Currently, there are 3,174 judicial posts in England, spread out widely across the country (Ministry of Justice, 2020). However, this is a small number when compared to the number of lawyers in England.

The judge's role emerges most clearly when parties litigate and go to trial. Generally, trials have two stages. The first is fact-finding, when both sides present their version of the facts to the court, and when the judge must decide which version is

---

[1] Some lawyers are employed by the government. For example, public defendants' salaries are paid by the state. However, their obligation in court is to their clients rather than to their employer. Finally, there are quasi-political legal roles, such as that of Attorney Generals in the US, which blur the distinction between a lawyer and a judge.

correct. A variety of different sources of evidence are used, such as documents, fingerprints, and witnesses (*inter alia* Wacks, 2015). In common law countries, the fact-finding role can also be carried out by randomly selected members of the public called the jury (Zweigert and Kötz, 1992).[2]

Second, after the facts are determined, the judge determines the outcome of the case by applying the law. Lawyers on both sides present their arguments for what the law is and how it fits the facts. Of course, they do so in a way that will benefit their side. The judge's duty is then to make a decision and explain it. Crucially, the judge is constrained in their decision. They are not free to consider the applicability of any law; instead, they may only consider the validity of the alleged violations presented to them. If the judgment is considered incorrect, it can be appealed. In this case, it goes to a higher court for review (Wacks, 2015).[3]

In common law countries, the judge has a law-making role as well. Through the doctrine of legal precedent, when a common law judge decides a case, they create a new legal rule. Future cases with similar facts to an already decided case must be adjudicated by judges in the same way (Black, 2019). In the civil law system, in contrast, past cases are not binding, but under the principle of consistent law application, they still carry a certain weight when new cases are decided.

2.2 The Rule of Law

The Rule of Law is a fundamental political principle recognized by both common and continental jurisdictions, underpinning how the legal system operates. Many different principles fall under the umbrella of the Rule of Law. We will focus on the following four.

Consistency. It is a basic principle that like cases should be treated alike, and that judicial biases should not interfere with legal decisions (Fuller, 1969). This is not an absolute principle, given that the law sometimes needs to develop to meet new societal needs, but it ensures the law

is predictable and consistent across cases. It is especially important for lower-tier judges, who are expected to dutifully apply the law set out by the courts above, such as the Court of Appeal and Supreme Court.

Access to Justice. Access to justice is another fundamental principle of the legal system (Diver, 2020). Legal subjects must be able to gain access to legal advice and have time in court to enforce their legal rights. Procedural delay, extremely high costs, and geographically sparse courts are all hindrances to access to justice and undermine the rule of law.

Equality Before the Law. A central principle of the modern liberal state is that no one is above the law, including lawmakers (Dicey, 1979). This means politicians and judges are subject to the rules they make. Equality in this way improves the legitimacy of law-making: Legal subjects can be assured that there is one legal regime, equally applied to them and the law-maker. The principle also acts as a checks-and-balances mechanism for lawmakers: They feel the sting of unfair and unjust legal rules because they are subject to them. Further, they can observe firsthand how the rules operate in practice, gaining useful feedback for creating new laws (Dicey, 1979).

Comprehensibility. Finally, it is an essential principle that legal subjects can access and understand the law which governs them (Fuller, 1969). This means the law must be publicly available, understandable, and not contradictory. It also requires satisfactory explanations for why legal decisions were reached. This makes it possible for legal subjects to follow the law, as well as giving them the ability to challenge and criticize legal rules (Hart, 1961; Raz, 1979).

2.3 Substantive Justice

The Rule of Law relates to the form rather than the content of the law. The latter is a matter of substantive justice, which concerns the question of whether the content of legal rules is morally good or bad. Law is connected to morality for three reasons (Green and Adams, 2019).[4] First,

---

[2]The prevalence of the jury differs between common law countries: In England, for example, they are present in many criminal cases but few civil cases (cases involving contracts, torts, conveyances). In the US, on the other hand, civil juries are more common in civil cases.

[3]When this occurs, it is typically only the legal arguments that are reconsidered. The facts are assumed to be what was decided in the first trial.

[4]There is a debate between legal positivists and natural lawyers on whether law *must* be substantively just to qualify as law (Gardner, 2001).

laws deal with inherently moral issues, such as abortion, homicide, and constitutional conflicts. Second, laws are often created to pursue moral ends, such as reducing crime, furthering racial and gender equality, and redistributing wealth in society. Third, law is a tool used by the state to coordinate itself, and as such can be used for both good and evil at a larger scale than would otherwise be possible.

Given these connections between law and morality, it is important to ensure lawmakers are accountable to those they govern. A morally acceptable legal system must have mechanisms of political accountability for lawmakers (Dicey, 1979). For example, in democratic countries, legislators are held accountable to the voting populace. As an extreme case, in the US judges are elected, much like politicians.

Overall, substantive justice is more difficult to assess than the Rule of Law principles. First, its content is inherently contestable. What constitutes a morally acceptable law? Who should be responsible for holding lawmakers to account? These issues are often socially and politically divisive and constitute matters that reasonable minds can differ on. Second, what is morally good changes over time. What was previously morally acceptable may become unacceptable and vice-versa. A good lawmaker must take these changes into account when making law to meet the evolving needs of society.

3 Automation from a Policy Viewpoint

A number of proposals exist on the future role of legal NLP (Susskind, 2008; Alarie et al., 2016; Alarie, 2016; Casey and Niblett, 2016, 2021; Goldsworthy, 2019). We taxonomize these proposals into two groups: (1) Those that advocate for the replacement of legal actors with technology, and (2) those that advocate for mere augmentation. The former, more radical view suggests NLP can be used to replace all legal tasks. The latter, on the other hand, proposes legal NLP can only supplement the work of traditional human lawyers and judges, without completely replacing them.

3.1 Replacement

The most ambitious proponents of legal NLP suggest the entire legal profession can be completely automated (Cobbe, 2020). There are broadly two approaches to replacement. The first suggests that a large amount of data, better models, and more computational power will allow the creation of highly effective legal NLP models (Alarie, 2016). Under this view, the models will eventually be able to give the correct answer to any legal question instantly, with the benefit of more information than any human lawyer or judge could ever possibly consider (Goldsworthy, 2019). The decisions of this machine lawyer would not be directly contestable by litigants or the accused, removing the need for courts and lawyers.

Under the softer replacement approach, while all judges and lawyers are replaced, a human remains in the loop as the policymaker (Alarie, 2016). A small number of human policymakers would set general policy objectives, such as to reduce traffic accidents by 40%. Legal NLP models with access to vast quantities of data then design and implement rules to achieve these policy goals (Casey and Niblett, 2016). As with the first approach, these rules would be set automatically by the legal NLP model and there would be few, if any, opportunities for appeal or adjudication. Without adjudication, the need for lawyers is greatly diminished (Casey and Niblett, 2021). Instead of hiring a lawyer and going to court, demands for change would have to be directed towards governmental policymakers.

3.2 Augmentation

More conservative Legal NLP tools of this variety are already used widely by legal professionals, in particular when carrying out due diligence, document review, regulatory compliance, and e-discovery. In 2018 in the UK, for example, 48% of law firms were using AI in some form (Walters, 2018). Tech startups cater to this extensive use of legal NLP and argue its use can reduce the number of lawyers needed for a particular task (Frey and Osborne, 2013). One proposal is to use legal question-answering systems to reduce the cost of legal advice, and thus make it far more available. Such technology could have important implications for access to justice, allowing clients to obtain legal advice that they might otherwise not be able to afford (Pasquale, 2019).

Another proposal is to use AI to reduce uncertainty in estimating the outcome of potential litigation. Firms have already begun to use software to predict the likely outcomes of a lawsuit; they could use this to advise a client about their

risks and liabilities (Ellen Gregg, 2019). On the public side, several governments have started testing the augmentation of judicial functions, such as sentencing and lower-tier tribunal decision-making in criminal trials. The company Equivant, for example, offers an AI product which predicts the re-offending rates of different criminals based on various characteristics. It is now used to make parole and detention decisions (Hildebrandt, 2020a). Controversially, automated judges already sentence people in the US (Kehl and Kessler, 2017) and China (Stern et al., 2020).

Finally, some have proposed that augmentation could be implemented in the hierarchy of the appellate court system (Cohen et al., 2023). Under this view, lower-tier judges, such as county court judges in England, could be replaced with automated decision-making software, while higher-tier judges would be retained and would review decisions from the automated decision-making tribunals under an appeal procedure. If reviewing judgments is indeed cheaper and faster than producing them, a legal expert could review each decision whilst still retaining many of the benefits of automation. Such a hybrid system would have the advantage of speed in the lower instances of the court, but ensure that legal precedent continues to be developed by experienced human judges.

4 Legal NLP

Legal NLP can trace its origins all the way back to the late 1950 (Kort, 1957; Nagel, 1963; Lawlor, 1963).[5] Hypo (Ashley, 1988), one of the earliest symbolic AI systems, explicitly encoded the principles of case-based reasoning in terms of analogy and difference, often based on hand-extracted features. Many others were inspired by it (Aleven and Ashley, 1997; Rissland and Skalak, 1991; Branting, 1991).

Aleven (2003) originally introduced the task of predicting an outcome of a case, which has nowadays become the modern benchmark of legal NLP. For a long time, Issue Based Prediction (Ashley and Brüninghaus, 2009), one of the successors of the Hypo model, held the state of the art for this task. However, the wider adaption of the symbolic systems was hindered by their reliance on humans

for both processing the input and encoding the changing rules of law. This became a problem for the deployment of these models since the law can change constantly with every new court decision.

Recently, the popularity of machine learning, combined with a desire for more robust models of law, has rejuvenated interest in developing applications for the legal domain. This shift from symbolic to sub-symbolic legal AI has implications for contemporary legal NLP research. The aim is now to approximate legal reasoning from legal data, rather than encoding it by hand. Since the architectures powering many SOTA approaches in legal NLP are fairly homogeneous—they are all pre-trained Transformer-based LLMs—the deciding factor is the choice of data the ML models are fine-tuned on and the number of parameters.

Legal ML tasks include question answering (Monroy et al., 2009), legal entity recognition (Cardellino et al., 2017), text summarization (Hachey and Grover, 2006), outcome prediction (Xu et al., 2020a; Zhong et al., 2018; Aletras et al., 2016), majority opinion prediction (Valvoda et al., 2018), legal topic classification (Nallapati and Manning, 2008), court opinion generation (Ye et al., 2018), case citation resolution (Shaffer and Mayhew, 2019), study of legal precedent (Valvoda et al., 2021), or applications in legal consulting (Wang et al., 2019). For a comprehensive overview of legal NLP, see Zhong et al. (2020a).

Yet another strain of work has emerged where the goal is to model individual judges using sociodemographic and similar features, instead of legal text. An example for this line of work is that of Katz et al. (2017), who predict the outcome of the US Supreme Court case law. The features used in this work are meta-data of the court, such as date, court position in the court hierarchy, judge names and lower court outcome; with these, a simple nearest neighbour classifier achieved 70% accuracy.

5 Ethical Dimensions of Legal NLP

We believe there are both benefits and risks in the use of NLP in the legal domain. The benefits are large enough that refusing to use NLP could present a moral failure akin to not using new medical treatments to treat patients. On the other hand, the ethical risks of NLP are also significant and need to be weighed up against these potential benefits.

---

[5]The field has gone by different names at different times. Juris-informatics, legal informatics, and legal artificial intelligence are just a few of these.

5.1 Benefits of Legal NLP

We see three major benefits that legal NLP can bring to the rule of law and the substantive content of the law. We consider each in turn below.

Accessibility. Technology could improve access to justice in the following three ways. First, automated legal services could be created and delivered more quickly than human advice, enabling the fast resolution of legal disputes. Second, what prevents access to the law for many citizens is not only the price of legal services, but also the impossibility to predict the cost at the start of a case. Both factors could significantly decrease if aspects of legal reasoning are automated. Third, legal NLP could facilitate geographically wider availability of legal services. Clients would no longer be constrained by the physical location of the providers of legal services. During the pandemic, when many courts adopted a hybrid or fully online operation for some of their hearings, this aspect of geographical freedom proved beneficial for many (Legg, 2021).

Consistency. There is a high variance in the quality of human-provided legal services. In practice, this means that different people have access to different quality of legal advice, and related cases might not be decided in similar ways. Legal NLP could help play a role in narrowing the gap between different legal actors, bringing us closer to the ideal of an equal and consistent legal system.

Capacity. NLP could help lawyers, judges, and litigants deal with growing legal complexity. Over time, more legal sources are produced, at a greater rate, in more detail, and with greater degrees of legitimacy. Lawyers have found this growing mass of material difficult to handle (Hildebrandt, 2018), and it is likely that the quantity and complexity of legal work will only increase in the future. Legal NLP could help to reduce this complexity and enable deeper and wider research than any human alone would be capable of (Alarie, 2016). For instance, the capacity of large language models, such as GPT-4 (OpenAI, 2023), to digest a scale of data beyond human comprehension might help to address this problem in the future.

5.2 Risks of Legal NLP

In addition to benefits, there are also important ethical risks in using NLP in the legal domain.

There are **technical** as well as **inherent** challenges. Current legal NLP technology encounter technical limits in achieving what skilled lawyers and judges can do (Clayton and Boyd, 2020). These are, in theory, resolvable with better models and improvements in the capacities of NLP. Inherent challenges, in contrast, are of an ethical and political nature; these remain problematic even if NLP were able to perfectly perform legal tasks at human level.

5.2.1 Technical Challenges

We discuss three technical challenges of legal NLP.

① **The Trial Problem.** As noted in §2, a central part of a trial is the fact-finding stage. Machines can comb through large quantities of textual data at speed, and indeed, current NLP techniques can easily deal with vast amounts of data. The nature of some types of evidence, however, excludes NLP tools from detecting and using it. The legally relevant factors in many cases are mental states, for instance intent, foresight, and malice, which humans try to detect when observing the defendant during the trial. Human judges have to assess the psychological characteristics of defendants or plaintiffs, such as their overall credibility and honesty, and whether they seem violent, reckless, or regretful. Trying to determine automatically if these mental states are present is a highly nuanced process, which cannot be done well from text alone (Tortora et al., 2020; Petoft and Abbasi, 2020). In fact, it is a hard task even for human judges and lawyers, who have the advantage of being physically present during the trial (Wistrich and Rachlinski, 2017).

To be able to use NLP models in court the models will need to uncover the true meaning of the situation at hand. The lack of grounded understanding is a well-known issue in the wider NLP discourse. In a reenactment of the Chinese room argument (Searle, 1980), Bender and Koller (2020) discuss the limitations of NLP models in accessing meaning from text alone with the introduction of their *octopus test*. Under the octopus test, two people stranded on separate islands communicate with a telegraph. Unbeknownst to them, a hyperintelligent octopus listens in on their conversations by monitoring the telegraph's underwater cable. Over time, the octopus can learn to predict the individual responses between the

two people well, simply by learning the statistical patterns of the language. However, Bender and Koller assert the octopus will never be able to truly understand the meaning of the conversation because it has never experienced the world on land. In other words, its understanding of the language is not *grounded* in meaning. Under this view, a language model will never be able to know a language to the degree a human does.

② **The Moral Dimension of Law.** Creating substantively just law requires a deep understanding of morality, politics, and changing social conditions. A system that has been trained on past legal material, as all current ML-based NLP models are, is likely to be backward-looking. It will therefore not be able to take an active part in evolving the law to meet the needs of society (Markou and Deakin, 2020). Leins et al. (2020) warnz that ML models will always have a tendency to lag behind the latest developments of legal doctrine. This leads Leins et al. to question the ethics of collecting legal data in the first place. In particular, they are concerned about the inability of updating the datasets when legal decisions get reversed or when a case is appealed to a higher court.

In common law systems, the legal system has a fluid aspect, and an important judicial role is re-interpreting existing precedents in light of current social needs (Delacroix, 2022). For example, the highly controversial case of *Roe v Wade* required the US Supreme Court to combine a mixture of modern moral and political considerations with the technical legal interpretation of the Constitution. The case was made complex not only by the difficult moral debate over abortion, and the political question of appropriate federal state power, but also legal questions over the correct way to interpret the Constitution in a changing society. As Talat et al. (2022) and Fraser et al. (2022) point out, there are problems with modeling morality with NLP methods. Current NLP models of morality only learn about moral situations from a limited context, which restricts their capacity to make complex and important moral decisions. Furthermore, in the wider NLP discourse, there is an ongoing discussion about how biases in the datasets used to train NLP models can get exaggerated by the models (Bender et al., 2021). A legal NLP tool acting upon such biases would have severe repercussions in the legal domain, particularly if such a tool is employed with the goal of upholding moral values.

③ **The Justificatory Role of Law.** Legal decisions are the product of the debate between two parties, in a matter that is often highly contentious. The resolution of this debate is important for the public and must therefore be *comprehensible* and *justifiable* to them. Current legal NLP models are unable to give reasons for their decisions. This weakens their ability to convince the parties to the dispute that a correct decision has been reached. Lack of explainability also goes against the principle of contestability of any legal decision, a central part of the rule that everyday citizens should be able to participate in, understand, and challenge the decisions which govern them. A legal system that cannot explain the outcomes of cases could severely restrict this possibility (Hildebrandt, 2020b).

The issue of contestability of an ML system goes beyond legal applications; it has been discussed in the wider NLP literature (Mitra, 2021), as well as in the context of human–computer interaction (Hirsch et al., 2017) and social computing (Vaccaro et al., 2019). There are also general limitations and privacy concerns regarding the use of LLMs, which we will discuss separately in §8. NLP has made huge gains in the past decades, and it hopefully will continue to do so. As NLP advances, we expect the technical challenges to get gradually resolved or at least ameliorated. But even if they do, we still have to contend with the inherent challenges.

5.2.2 Inherent Challenges

We foresee three inherent challenges in using technology in the legal domain.

① **Centralized Power.** The benefits of using NLP in the legal system, be it accessibility, consistency, or capacity, lie in the scalability of the technology. The underlying assumption is that with technology, there will be fewer human legal actors involved in any single legal process. The flip side of these benefits, however, is the risk of power being centralized. Without AI, the expertise required to adjudicate legal cases is possessed by judges and lawyers (Cobbe, 2020), and this expertise is held in a distributed manner. These actors work separately from one another: most of them are not coordinated in one group or location

at any time. They can therefore act as checks on one another. Replacing and augmenting legal expertise with technology will increase efficiency, but is likely to result in a situation where there are fewer individual actors in the system. This creates the risk of upsetting the current fine-tuned balance, as fewer legal agents would operate with fewer checks.

② **Increased Brittleness.** The risk of centralizing the legal system also raises the danger of increased fragility. A legal system dependent on digital infrastructure is more prone to technical failures such as electrical outages or cyber-attacks. The risk of widespread system failure increases as legal NLP proliferates the legal system. If a social system cannot accommodate unusual or extreme events, it has been described as fragile (Taleb, 2012).

Another source of fragility could come from the speed at which a legal system operates. The expected speed-up of the entire legal process is of course a major benefit of successful automation. A fast legal system will also result in the fast creation and application of new precedents. But this also reduces the margin of error in law-making, as there is less time to review, challenge, and test new law before it is re-applied. With fewer checks at the application stage, unworkable or unjust laws could be applied, and they could be applied more quickly and comprehensively. In areas where predictability is crucial, such as land law and contract law, small errors could be devastating if they were applied across the legal system. Therefore, decisions must be reached with a high level of reliability.

③ **Lack of Accountability.** Legal actors must always be held accountable for their decisions, following from the legal principle of equality before the law (§2.2). Currently, legal decisions are made by identifiable human experts, which ensures that they can be challenged for their decisions, both by members of the public and by private clients (Diver, 2020). One risk of introducing legal NLP into the system is that it makes it more difficult to identify who is responsible for specific decisions. The problem of accountability has been raised with respect to automation at large and in particular in the context of self-driving cars and recidivism prediction tools (Gless et al., 2016; Tonry, 2014; Ryan, 2020).

When it comes to partial automation, attribution of liability is particularly complex, as the case of self-driving cars shows. The liability for self-driving cars typically depends on the level of automation and user control (Ryan, 2020; Boeglin, 2015). SAE defines six levels of automation for self-driving cars.[6] Level 0 means no automation. Level 1 (Driver Assistance) and Level 2 (Partial Automation) describe a vehicle with the ability to support the driver in steering, braking, and accelerating. Level 4 (High Automation) is where the car is capable of fully driving itself in some driving modes, for example on a highway. Level 3 (Conditional Automation) is in between: The car is somewhat self-driving, but a human must be ready to intervene at any point. Level 5 is a fully autonomous car.

Manufacturers will usually deny responsibility at the lower levels (Level 1 and Level 2). Conversely, they will take responsibility for crashes occurring at Level 4. At Level 3, there is considerable uncertainty: Sometimes the driver is found liable for reduced damages, whereas in other cases they escape liability (Ryan, 2020). The above taxonomy, when transferred to the legal NLP case, might be a useful method of assessing where the pitfalls of deploying legal NLP might lie. A Level 3 NLP legal system would be one where automation of an aspect of legal reasoning takes place, but humans are expected to monitor the system for potential failures.

Another lesson learned comes from recidivism prediction tools, namely, about the difficulty of challenging legal decisions aided by software (Thomas and Ponton-Nunez, 2022). In the US, algorithmic risk assessment tools have been employed to aid judges in making predictions about the risk of the defendant re-offending before the trial.[7] Since the deployment of these systems, the validity of their use in the court has been challenged several times by applicants alleging an infringement of their constitutional rights.

For example, in the case *State v. Loomis*, the accused claimed that the use of risk assessment

---

[6]The SAE Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems breaks down into six levels (0-5) and can be found at here.

[7]For example, Correctional Offender Management Profiling for Alternative Sanctions, the Historical Clinical Risk Management-20, the Violent Risk Assessment Guide-Revised and Sexual Violence Risk-20.

tools to calculate his criminal sentence breached his rights. One of his arguments was that the software, due to its proprietary nature, was unaccountable and precluded him from challenging its scientific validity. The court rejected this argument on the grounds that, although the algorithm was hidden, the data it used was publicly available in its entirety. Worryingly, the courts presiding over these cases have not found the lack of transparency as an infringement to the right of the parties involved (Thomas and Ponton-Nunez, 2022). Their reasoning relies upon the fact that, much like in the case of self-driving cars, where automation is only at Level 1 or 2, there is an identifiable human (the trial judge) who is primarily responsible for the relevant decisions.

If these early cases are anything to go by, challenging a machine-made decision on the basis of accountability will be difficult, especially where such decisions are made using proprietary technology. Further, such challenges are apt to get more complicated if automation is employed in more situations: resolving who should take responsibility in such a situation will require inherently political choices (Ryan, 2020). Should it be the responsibility of the researchers who have designed the legal NLP system, the software engineers who have implemented it for the provider, or the provider of the service (Morison, 2020)?

As a matter of equality before the law, it is also important for those responsible for the decisions to feel the effect of the laws they create. System administrators are far less exposed than trial judges to the specific decisions and outcomes their tools produce. The questions of accountability in legal NLP, therefore, go beyond technical concerns.

6 Judge vs. Lawyer

Since lawyers and judges both practice law, the use of technology to automate aspects of their roles should give the public wider access to more consistent and better-informed legal services, and should overall benefit the accessibility, consistency and capacity of the legal system. Both the automation of the lawyer and the judge would result in these positive effects. We will now discuss the roles with respect to the risks, and whether they are also equally balanced between the two primary legal actors.

6.1 Technical Challenges: Judge vs. Lawyer

We will first compare the automation of judges and lawyers in the face of technical challenges.

① **The Trial Problem.** The judge and the lawyer have different roles during the trial. The judge is responsible for assessing witness testimonies, a process where they examine inconsistencies, reactions, and emotions from the oral accounts of witnesses and litigants. The judge, therefore, must be able to read diverse physical cues and contextual social knowledge, skills that are difficult to replicate through textual learning. The role of the lawyer, on the other hand, is to ascertain their client's side of the story, assess their opposing counsel's evidence, communicate with clients and processing ambiguous facts. They also need to assess the changing reactions of a jury to arguments. Lawyers also require the skills to cross-examine witnesses and create compelling narratives, tailored for the particular judge or jury composition (Brooks and Gewirtz, 1996). All of this requires emotional and social sensitivity. Therefore, although their roles are different, lawyers and judges face similar difficulties when it comes to assessing standard legal evidence.

Recommendation 1 (Multi-modality) *Legal NLP needs multi-modal approaches, integrating vision and sound, which are necessary for the full automation of judges and lawyers.*

② **The Moral Dimension of Law.** The requirement of substantive justice in law means lawyers and judges inevitably need to consider ethical and political factors. Many cases involve morally charged issues, and judges are under a duty to the public to reach morally correct decisions. We believe that the role of a judge as a moral arbiter sets a high burden for the full automation of judges. The metric for morally acceptable law is highly contestable and controversial. A successful judgment under this criteria must be persuasive to a wide range of groups, ranging from litigants, politicians, academic commentators, and the public generally. Given the controversial nature of moral judgments, these groups are likely to diverge, and deciding the correct weight to be accorded to each is also contestable. This makes it extremely difficult to determine whether a model is successful in making morally good decisions.

In contrast, a lawyer only needs to successfully appeal to the decision-making of the judge rather than create an argument they themselves believe to be morally sound. It is not their primary responsibility to ensure law is enforced according to moral standards accepted broadly by different groups in society.[8] Rather, their role is to present their client's case in the strongest, most persuasive form possible to convince the judge to decide in their favor. A good argument, under this standard, is one which generally persuades judges and secures victories in lawsuits for the client. Lawyer automation has therefore a lower technical burden to overcome when it comes to modeling morality.

Recommendation 2 (Morality) *In line with Talat et al. (2022), we suggest including domain experts in developing the moral sense of NLP models.*

③ **The Justificatory Role of Law.** Both lawyers and judges must explain their decisions. However, their justifications serve very different ends. The rule of law principle of comprehensibility incentivizes judges to explain their decisions. It is important for the members of the public to understand why a judgement was reached, both for its legitimacy, and also so that they can contest it. Justifying decisions to the public to this level sets a very high technical burden to overcome. Opinions can differ on whether an explanation is comprehensible, which means that feedback from a wide range of public actors and interest groups is required. Furthermore, the assessment of the quality of the judge's reasoning and whether it has shaped the law in a positive or negative way can only be assessed by longitudinal studies, if at all. Justifications given by lawyers, on the other hand, are often of a different, more practical nature. The main test of the quality of a justification is, again, only whether their arguments can persuade judges and other lawyers, which is much easier to measure. If other lawyers and judges accept and understand the arguments, the justification was successful; if not, it requires improvement. From the viewpoint of justification, it is therefore also the case that the burden of automating a judge is higher than that of automating a lawyer.

---

[8]That does not mean that lawyers don't talk about ethics in court. Since judges take moral factors into account, lawyers must also include them in their arguments, by indicating the ethical context of their client's situation (Liu, 2022).

Recommendation 3 (Explainability) *Legal NLP models need to explain their reasoning in a way that allows members of the public to exercise moral and political scrutiny over it.*

To summarize, there are outstanding technical challenges for the full automation of the judge or lawyer. In light of these challenges, we believe that augmentation is a more realistic pursuit given the current limitations of legal NLP. We believe that aspects that should be left aside from automation are witness assessment, moral decision-making and justifications to human experts. Nonetheless, we see the pursuit of the full or partial automation of lawyers as less problematic than that of judges.

6.2 Inherent Challenges: Judge vs. Lawyer

Now we turn to the challenges that cannot be resolved by technology alone.

① **Centralized Power.** The risk of centralizing power is much greater if judges are automated than if lawyers are automated. There are two reasons for this: The political power judges possess, and the organizational structure. First, as noted in §2, judges in common law countries possess law-making power, which gives them influence over the lives of everyday people. The role of a lawyer is different. Lawyers do not create new laws, they only make arguments that the judges take into account. This considerably limits their ability to change the law. Second, the organizational structure centralizes judges' power to a far higher degree than lawyers' power. As state employees, judges are by default employed by a single organization: the government. Additionally, the role of the judge involves high consistency. This fact has, in contrast to our main argument here, tempted some to believe that a single automated decision-making system is desirable in the first place, see §3.

Conversely, lawyers are organized in private firms which compete with one another for the business of clients. Law firms are free to structure themselves and tailor the advice they give to their clients. As a result, law firms are more likely to operate independently, choosing their own tools and approaches, whether they are aided by technology or not. Consequently, it is unlikely that the automation or augmentation of the lawyer

profession would be achieved using a single NLP model or tool, and therefore the risk of centralizing power is smaller when it comes to the role of lawyers in the judicial system.

Nonetheless, the risk of lawyer centralization is not zero. Legal NLP, like other technology sectors, might demonstrate network effects and lock-in, leading to one or two providers dominating the market. This risk, connected with the potential effect on the development of the law and the accessibility of legal advice, means centralization remains a concern when automating either actor.

② **Increased Brittleness.** Increased centralization also brings the risk of increased brittleness. Since judges are more at risk of the former, they are also more susceptible to the latter. The importance of the role of the judge as a lawmaker magnifies the potential harm that could arise from a failure or malfunction of the tools they use. While it is true that a human judge can be biased, incompetent, or subversive, it is also true that they can be checked by other judges. But this kind of self-policing among judges is threatened when aspects of the judicial system are handed over to machines. As discussed above, the speed and scale at which automated judicial reasoning could cause harm is much greater than in the case of a biased human. Some of these issues are mediated when we move to partial replacement, because human review is introduced. The risk decreases with more human oversight, subject to a trade-off between speed, cost, and fragility.

In contrast, lawyers have less influence on the system as a whole, and there is a lower risk of their centralization. Therefore, in the case of their replacement or augmentation, there is also a lower risk of brittleness. Clients who are able to choose from a variety of legal service providers can avoid bad actors; faults in tools used by one law firm is therefore unlikely to spill over into the others.

Recommendation 4 (Diversity) *It is necessary to maintain the diversity of adjudicators in order to avoid the risks of power concentration and brittleness. Whether machine or human, using the same automated system for every case poses a danger of systemic malfunction and fragility.*

③ **Lack of Accountability.** The judge and lawyer are both accountable, but they are accountable to different stakeholders. The judge is accountable to the public and state officials, whereas lawyers are accountable directly to their clients. This distinction has implications for the automation of lawyers and judges. Law firms must meet the needs of their clients; they are liable for malpractice and must deliver services which are competitive in the legal services market. As a result, they can naturally act as the site of responsibility should their use of legal NLP fail. In contrast, judges are accountable to the public and state officials. Unlike for lawyers, for judges there are no market effects that could evaluate the success of the legal NLP used to augment or replace them. In that case, it might become difficult to identify any individuals responsible for the decision made by a machine judge.

Even if the use of AI in the court is limited, accountability remains an issue. If current recidivism prediction software is anything to go by (see §5.2.2), it will be difficult to challenge any negative impact a statistically driven tool might have on the judge's decision. Currently, a claimant has no right to examine the tool that is used to help to decide their legal faith and can only hope that the judge's understanding of the technology is sophisticated enough to account for the biases and errors such a tool might introduce.

There are also issues with the principle of equality before the law. A machine judge which is used to automatically decide cases and, therefore, make new precedents, will have no experience of the law it enacts. A human judge experiences the effects of the cases they decide on, but a machine will lack the basic feedback process to make this possible. This raises issues of the legitimacy of the decisions of machine judges. Since lawyers do not create law, their legal NLP replacement does not pose this risk.

Recommendation 5 (Accountability) *It is necessary to decide who is accountable for the actions made using legal NLP.*

Resolving the inherent challenges requires careful consideration of how the technology should be introduced in the courtroom. While the hybrid augmentation approach seems the most sensible, a danger remains that the human in the loop will overestimate the abilities of the NLP tools and defer to them. This is particularly concerning when it comes to the automation of the role of a judge.

7 The Voice of a Judge in NLP

In the above discussion, we argued that the automation of the lawyer poses fewer challenges than the automation of the judge. Nonetheless, the tasks lawyers and judges conduct during their work are closely connected. In this section, we explore several legal NLP tasks and demonstrate how the role of a judge is inadvertently modeled via reliance on judge-created training data. We then make suggestions on how these tasks could be redefined to prioritize the voice of a lawyer.

7.1 Legal NLP Tasks

The three tasks we consider in this section are: Legal Outcome Prediction, Similar Case Matching, and Legal Question Answering. Each task relies on judge-generated text for training data. Since the major paradigm in legal NLP is ML, the current operationalizations of these tasks turn them into a series of different approaches for modeling a judge. It is worth noting from the outset that these tasks and their accompanying datasets are not an exclusive list of legal NLP research directions; see §4. Nonetheless, these are tasks used in well-established legal benchmarks, such as LexGLUE and COLIEE (Chalkidis et al., 2022; Rabelo et al., 2020), and thus deserve our attention.

The first task we consider is that of **Legal Outcome Prediction**. Given the facts of a case, the task is to predict the outcome of a case (Chalkidis et al., 2019). While both judges and lawyers are interested in estimating the potential of a case to succeed, they approach this problem from different angles. As noted above, lawyers are interested in maximizing the chance of winning the case; they do so through the arguments they create for their clients. Judges, on the other hand, are interested in establishing a sound precedent and serving justice. They base their decisions on the situation of the claimant, the arguments presented by the lawyers, and their understanding of the law. The legal NLP models trained for this task do not correspond to either role above. Instead, both inputs and outputs the models are trained on are extracted from *cases*.

A case is a transcript of the judge's reasoning towards the outcome. To better understand why the ML operalization of the task is problematic, let's first have a closer look at the model inputs, the facts. When judges describe the facts of a case,

they have already decided on the case outcome. Therefore, the facts used as input to the models are part of the judge's argument. A real judge has access to all the evidence presented by the parties to the court. Legal NLP datasets do not currently contain this information. Therefore, Legal Outcome Prediction models decide on the outcome given only a small subset of highly curated data when compared to a judge. Because the fact selection process introduces clues about the outcome of a case, the task of predicting the outcome from this data becomes artificially easy, at least when compared to a situation where the input instead consists of an unbiased full description of what had happened. Furthermore, from a legal standpoint, relying on facts alone is insufficient because the case decision is never made solely on facts. The reliance on only the facts as an input to the model turns Outcome Prediction into an artificial task, one that would make sense only if we replaced both legal actors with AI. We have argued in the previous section why this is undesirable.

Now we can turn to the model outputs; the outcomes. The most popular legal NLP treatment is to cast the task as a binary decision over all possible laws. The idea is to predict which laws have been violated, given the facts of a case. Formally, the model is trained to predict a binary vector $\{1,0\}^K$, where 1 represents a violation of one of the K laws under consideration. But it is not the case that a judge passes a decision over every law there is. Instead, a judge makes a decision with respect to a subset of laws, namely, those the lawyer has alleged as violated. We will refer to these as claims from now on. If no knowledge about claims is taken into account, 0 is left ambiguous: It can represent either a law that has been claimed as violated but the judge has decided it is not, or a law that is completely unrelated to the facts at hand. As Valvoda et al. (2023) point out, the standard operalization of the task is therefore artificially easy.

The second task is **Similar Case Matching**. Under the COLIEE competition Task 1 (Rabelo et al., 2020), given a text of a case with redacted references to previous case law (i.e., the precedent), the task is to correctly predict the redacted references (but not their location in the original text), see Table 1.[9] While both lawyers and judges

---

[9]Not all datasets for similar case matching have been sourced this way. Some, like Xiao et al. (2019), have been manually created and do not fall in the scope of our critique.

Legal Outcome Prediction - Chalkidis et al.

Facts: *''Ms Ivana Dvořáčková was born in 1981 with Down Syndrome (trisomy 21) and a damaged heart and lungs. She was in the care of a specialised health institution in Bratislava. In 1986 she was examined in the Centre of Paediatric Cardiology in Prague-Motol where it was established that...''* for more see: Case of Dvoracek and Dvorackova v. Slovakia

Outcome: Articles: 2, 6

Similar Case Matching - Rabelo et al.

Query: *''The Plaintiff stated that, on the evening of the incident, he was in the telephone area waiting to use the phone. The assailant jumped the queue in an attempt to use the phone. The Plaintiff and the assailant ''bumped shoulders''...''*

Precedent: 010, 151

Legal Question Answering - Zhong et al.

Question: *''What crimes did Alice and Bob commit if they transported more than 1.5 million yuan of counterfeit currency from abroad to China?''*

Answer: Smuggling counterfeit money.

Table 1: Examples for three popular legal NLP tasks.

search for relevant precedents, the data used to train the models to find related cases comes solely from the judge, as was the case in the outcome prediction task. However, it is not the case that judges select these cases under some objective metric of relatedness; instead, they cite cases that support their arguments towards the outcome of the case.

This raises two issues. First, one can safely presume that, much like the facts above, the judge-cited cases are an incomplete set of relevant case law. They almost certainly contain only a portion of the citations that were used by the two parties to the case. In particular, they are likely to lack some of the precedent that was relied upon by the party that has lost the case. After all, if the judge agreed with the losing party's argument, she would not decide against it. This means that by relying on the precedent selected by a judge,

the task favors the view of a judge. However, the precedents that the lawyers have relied on are equally an indicator of relatedness between any two cases. These precedents are currently not captured by the COLIEE dataset.

Second, the precedent prediction task is inherently connected to the outcome prediction task above. If a lawyer wants to claim their client is innocent, they will be looking at a very different legal argument than if they were to claim that their client is guilty. By ignoring what role in the argument the precedent played, a case retrieved by a Similar Case Prediction model can either support the desired outcome or be against it. From the perspective of a lawyer, this distinction is crucial when searching for case law to build their argument around.

Finally, we turn to the task of **Legal Question Answering**. At first blush, Legal Question Answering might seem like an emulation of a lawyer. After all, lawyers are paid to answer legal questions. However, a closer inspection reveals that the questions in the Legal-Domain Question Answering Dataset (Zhong et al., 2020b) are predominantly about the inference of the crime committed, rather than the explanation of legal concepts. From the 26,365 questions in the dataset, 16,604 are case analysis questions, such as the one in Table 1. The primary difference from the Legal Outcome Prediction task is that the facts are stylized and simplified. Consider the Legal Outcome Prediction case from Table 1. If we compare it to the Legal Question Answering example below it, we can see that the tasks are very similar. In both instances, the goal is to predict an outcome of a case, given the facts. This is, again, an automation of the adjudicatory role of a judge.

7.2 Proposed Solutions

We propose two solutions to the above problems: utilizing the information about legal claims and collecting new datasets of legal briefs. Both claims and briefs are a product of a lawyer. Therefore, reconstructing the above tasks around datasets of lawyer-generated text turns them from tasks that model a judge into tasks that model a lawyer. Consider how this shift would be achieved by utilizing legal claims. For the Legal Outcome Prediction task and the related Legal Question Answering task, having the knowledge of both legal claims and outcomes can remove the ambiguity

inherent in the binary classification setting. The 0 category would be disambiguated into negative outcomes and null outcomes, allowing us to perform a three-way classification. Negative outcomes are the claims that have not succeeded in the court, null outcomes are unclaimed laws. In previous work (Valvoda et al., 2023), we have implemented such three-way classification and demonstrated significant improvements over previous approaches.

Alternatively, one could train the model to predict the claims directly: claim prediction can be reasonably defined as a binary classification task (Chalkidis et al., 2021). This is a simple solution since claim prediction is already the subject of LexGLUE Task B (Chalkidis et al., 2022). Claims are the product of a lawyer, therefore, the first approach incorporates the information about a lawyer in the legal outcome prediction task, while the latter approach directly models the role of a lawyer.

Let us now consider how collecting legal briefs could help. Legal briefs are the arguments the lawyers present to the judge on behalf of their client. They also contain evidence the argument is based on. A dataset of these briefs could address the limitations of facts as the sole inputs to the Legal Outcome Prediction and Legal Question Answering models. Specifically, having access to briefs would allow for training models conditioned on the full facts of the case contained in the legal briefs. Better yet, the model outcome could be conditioned on both the factual description of the situation at hand and the lawyers' arguments. By including the lawyers' arguments as input to the model, the outcome prediction task would stop implicitly assuming a fully automatic legal process, where a judge operates without interaction with a lawyer, but rather a process with two distinct legal actors: lawyers who are developing arguments and judges who are evaluating these arguments.

From a practical perspective, a lawyer could use such a model to estimate their chances of winning a case, given the arguments they develop. As for the Similar Case Matching task, legal briefs would allow for training models that consider the full spectrum of precedent relevant to a case. Opening access to this data, one could begin to develop legal NLP models that have the ability to find the precedent relevant to a desired outcome of a case.

In conclusion, the tasks described in this section are restrained by the lack of access to the data produced by a lawyer, which makes them model the role of a judge. The proposed solution is not difficult in practice. Legal claims are already extracted for the ECtHR dataset, so the shift is a matter of choice of the task we should prioritize when training our models.

8 Related Work

There is a growing field of research on NLP data privacy (Klymenko et al., 2022). Concerns over privacy naturally arise from training NLP models on sensitive data, such as medical records (Yadav et al., 2016). These concerns are relevant to legal NLP in particular, since the content of legal documents is often confidential and might additionally contain medical details. Since training data can be reconstructed from large language models (Carlini et al., 2021), which underpin much of the current research in NLP and legal NLP, data privacy is a pressing issue. One approach for privacy-preserving techniques for legal NLP is the differential privacy approach, a cryptography-based approach using transformer language models developed by Yin and Habernal (2022).

Going beyond privacy concerns, machine learning approaches to NLP open the systems up to ethical questions by training on human-generated data (Hovy and Spruit, 2016). Data from the internet, and from social media in particular, may contain biases which, if left unchecked, could adversely affect users on NLP systems trained on such polluted data. As far as legal NLP is concerned, LLMs harbor biases that can affect legal NLP models built on top of them. The legal data itself also contains biases which we would not want to be replicated in our models (Posner, 2008).

Some researchers have conducted studies on uncovering and quantifying different types of bias (Buolamwini and Gebru, 2018; Blodgett et al., 2020; Xu et al., 2020b; Stańczak et al., 2023; Maudslay et al., 2019). Others have started to develop tools to de-bias the models (Ravfogel et al., 2022; Bolukbasi et al., 2016; Belrose et al., 2023). With the scale of datasets growing over time, these issues are likely to grow (Bender et al., 2021).

Finally, while the law can be viewed as a codification of moral reasoning to some extent, very

little has been written about the ethics of automating the legal system. There has been an ongoing critique of NLP research aimed at automating legal sentencing and even calls not to publish this kind of research (Leins et al., 2020). In response to this criticism, others have warned against the threat of moralism impinging on academic freedoms (Tsarapatsanis and Aletras, 2021). In contrast to the previous work, we focus on the roles of the legal actors and the ethical implications of their automation.

9 Conclusion

In this work, we have evaluated the ethical and practical feasibility of different proposals for NLP research in law. By surveying a number of popular legal NLP tasks, we identify that the datasets they are built on favour the voice of the judge. However, the criteria for what makes a good judge include moral discretion and political accountability, and we therefore believe that practical applications of these kinds of NLP models face difficult challenges. For a machine judge to be successful, it would need to have capabilities far exceeding what is currently available. These capabilities include moral and social intuition; the sophisticated ability to give explanations; and, in turn, the ability to receive feedback from members of the public. Even if these technical challenges were met, there are further inherent risks to using legal NLP in the real world. We have discussed three of these: the centralization of power, brittleness and lack of accountability. Comparing the judge and a lawyer, we find the role of a lawyer less susceptible to these challenges. Our appeal is, therefore, to focus on the voice of the lawyer in future legal NLP datasets. We believe that overlooking the role of the lawyer hinders current neural approaches for modeling law. For the academic pursuit of furthering legal NLP research, too much focus on the text produced by judges hides the fruitful and interesting interplay between the lawyers that fuels the legal discourse.

Acknowledgments

References

Benjamin Alarie. 2016. The path of the law: Towards legal singularity. *University of Toronto Law Journal*, 66(4):443–455. https://doi.org/10.3138/UTLJ.4008

Benjamin Alarie, Anthony Niblett, and Albert Yoon. 2016. Regulation by machine. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2878950

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. https://doi.org/10.7717/peerj-cs.93

Vincent Aleven. 2003. Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. *Artificial Intelligence*, 150(1–2):183–237. https://doi.org/10.1016/S0004-3702(03)00105-X

Vincent Aleven and Kevin Ashley. 1997. Teaching case-based argumentation through a model and examples empirical evaluation of an intelligent learning environment. In *Artificial Intelligence in Education*, volume 39, pages 87–94. IOS Press.

Kevin Ashley. 1988. *Modelling Legal Argument: Reasoning with Cases and Hypotheticals*. Ph. D. thesis. Order No: GAX88-13198.

Kevin Ashley and Stefanie Brüninghaus. 2009. Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law*, 17(2):125–165. https://doi.org/10.1007/s10506-009-9077-9

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on*

Fairness, Accountability, and Transparency, pages 610–623. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Henry Black. 2019. *Black's Law Dictionary*, 11th edition. Thomson Reuters.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of ''bias'' in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Jack Boeglin. 2015. The costs of self-driving cars: Reconciling freedom and privacy with tort liability in autonomous vehicle regulation. *Yale Journal of Law and Technology*, 17:171–204.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356–4364. Curran Associates.

Karl Branting. 1991. Reasoning with portions of precedents. In *Proceedings of the 3rd international conference on Artificial intelligence and law*, pages 145–154. Association for Computing Machinery. https://doi.org/10.1145/112646.112664

P. Brooks and P. Gewirtz. 1996. *Law's Stories: Narrative and Rhetoric in the Law*. Law literature studies. Yale University Press.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In

Conference on Fairness, Accountability, and Transparency, pages 77–91. Proceedings of Machine Learning Research.

Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. 2017. Legal NERC with ontologies, Wikipedia and curriculum learning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 254–259. Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2041

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium*, pages 2633–2650. USENIX Association.

Anthony Casey and Anthony Niblett. 2016. Self-driving laws. *University of Toronto Law Journal*, 66:1–14. https://doi.org/10.3138/UTLJ.4006

Anthony Joseph Casey and Anthony Niblett. 2021. The death of rules and standards. *Algorithmic Regulation and Personalized Law*.

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. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1424

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. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.22

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. Lex-GLUE: 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. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.acl-long.297`

Jay Clayton and William Boyd. 2020. The second wave of algorithmic accountability. *Law & Political Economy Project*.

Jennifer Cobbe. 2020. *Legal Singularity and the Reflexivity of Law*, 1st edition, Is Law Computable? Critical Perspectives on Law and Artificial Intelligence, pages 107–134. Hart Publishing. `https://doi.org/10.5040/9781509937097.ch-005`

I. Glenn Cohen, Boris Babic, Sara Gerke, Qiong Xia, Theodoros Evgeniou, and Klaus Wertenbroch. 2023. How AI can learn from the law: Putting humans in the loop only on appeal. *NPJ Digital Medicine*, 6(1):160. `https://doi.org/10.1038/s41746-023-00906-8`, PubMed: 37626155

Sylvie Delacroix. 2022. Diachronic interpretability and machine learning systems. *Journal of Cross-disciplinary Research in Computational Law*, 1(2).

A. V. Dicey. 1979. *The Rule of Law: Its Nature and General Applications*, pages 183–205. Palgrave Macmillan. `https://doi.org/10.1007/978-1-349-17968-8_5`

Laurence Diver. 2020. Computational legalism and the affordance of delay in law. *Journal of Cross-disciplinary Research in Computational Law*, 1(1).

Daniel Smith Ellen Gregg, Bill Koch. 2019. How artificial intelligence is impacting litigators. *Attorneys' Liability Assurance Society Loss Prevention Journal*.

Kathleen C. Fraser, Svetlana Kiritchenko, and Esma Balkir. 2022. Does moral code have a moral code? Probing Delphi's moral philosophy. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing*, pages 26–42. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.trustnlp-1.3`

Carl Frey and Michael Osborne. 2013. The future of employment: How susceptible are jobs to computerisation? *Oxford Martin*, 114. `https://doi.org/10.1016/j.techfore.2016.08.019`

Lon L. Fuller. 1969. *The Morality of Law: Revised Edition*. Yale University Press.

John Gardner. 2001. Legal positivism: 5 and half myths. *American Journal of Jurisprudence*, 46. `https://doi.org/10.1093/acprof:oso/9780199695553.003.0002`

Sabine Gless, Emily Silverman, and Thomas Weigend. 2016. If robots cause harm, who is to blame? Self-driving cars and criminal liability. *New Criminal Law Review*, 19(3):412–436. `https://doi.org/10.1525/nclr.2016.19.3.412`

Daniel Goldsworthy. 2019. Dworkin's dream: Towards a singularity of law. *Alternative Law Journal*, 44(4):286–290. `https://doi.org/10.1177/1037969X19875825`

Leslie Green and Thomas Adams. 2019. *Legal Positivism*. Metaphysics Research Lab, Stanford University.

Ben Hachey and Claire Grover. 2006. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345. `https://doi.org/10.1007/s10506-007-9039-z`

Herbert Hart. 1961. *The Concept of Law*. Oxford University Press. `https://doi.org/10.1093/he/9780199644704.001.0001`

Mireille Hildebrandt. 2018. Law as computation in the era of artificial legal intelligence: Speaking law to the power of statistics. *University of Toronto Law Journal*, 68(1):12–35. `https://doi.org/10.3138/utlj.2017-0044`

Mireille Hildebrandt. 2020a. *Law for Computer Scientists and Other Folk*. Oxford University Press. `https://doi.org/10.1093/oso/9780198860877.001.0001`

Mireille Hildebrandt. 2020b. 'Legal by design' or 'Legal protection by design'? In *Law for*

Computer Scientists and Other Folk. Oxford University Press. https://doi.org/10.1093/oso/9780198860877.003.0010

Tad Hirsch, Kritzia Merced, Shrikanth S. Narayanan, Zac E. Imel, and David C. Atkins. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, pages 95–99. Association for Computing Machinery. https://doi.org/10.1145/3064663.3064703, PubMed: 28890949

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2096

Ministry of Justice. 2020. Diversity of the judiciary: Legal professions, new appointments and current post-holders. UK Government Statistics.

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). https://doi.org/10.1371/journal.pone.0174698, PubMed: 28403140

Danielle Kehl and Samuel Ari Kessler. 2017. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. Berkman Klein Center for Internet & Society, Harvard Law School.

Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential privacy in natural language processing the story so far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.privatenlp-1.1

Fred Kort. 1957. Predicting supreme court decisions mathematically: A quantitative analysis of the ''right to counsel'' cases. *American Political Science Review*, 51(1):1–12. https://doi.org/10.2307/1951767

Reed C. Lawlor. 1963. What computers can do: Analysis and prediction of judicial decisions. *American Bar Association Journal*, 49(4):337–344.

Michael Legg. 2021. The Covid-19 pandemic, the courts and online hearings: Maintaining open justice, procedural fairness and impartiality. *Federal Law Review*, 49(2):161–184. https://doi.org/10.1177/0067205X21993139

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. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.261

Sida Liu. 2022. *Between Rules and Power: Finding a Place for Lawyers in the Sociology of Professions*, 1 edition, Lawyers in 21st-Century Societies : Vol. 2: Comparisons and Theories, pages 445–460. Hart Publishing. https://doi.org/10.5040/9781509931248.ch-019

Christopher Markou and Simon Deakin. 2020. *Exploring the Limits of Legal Computability*, 1st edition, Is Law Computable?: Critical Perspectives on Law and Artificial Intelligence, pages 31–66. Hart Publishing. https://doi.org/10.5040/9781509937097.ch-002

Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5267–5275. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1530

Tanushree Mitra. 2021. Provocation: Contestability in large-scale interactive NLP systems. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 96–100. Association for Computational Linguistics.

Alfredo Monroy, Hiram Calvo, and Alexander Gelbukh. 2009. NLP for shallow question answering of legal documents using graphs. volume 10, pages 498–508. Computational Linguistics and Intelligent Text Processing. https://doi.org/10.1007/978-3-642-00382-0_40

John Morison. 2020. *Towards a Democratic Singularity? Algorithmic Governmentality, the Eradication of Politics; And the Possibility of Resistance*, 1st edition, Is Law Computable?: Critical Perspectives on Law and Artificial Intelligence, pages 85–106. Hart Publishing. https://doi.org/10.5040/9781509937097.ch-004

Stuart S. Nagel. 1963. Applying correlation analysis to case prediction. *Texas Law Review*, 42:1006.

Ramesh Nallapati and Christopher D. Manning. 2008. Legal docket classification: Where machine learning stumbles. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 438–446. Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613771

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Frank Pasquale. 2019. A rule of persons, not machines: The limits of legal automation. *George Washington Law Review*, 87(1):1–55.

Arian Petoft and Mahmoud Abbasi. 2020. Current limits of neurolaw: A brief overview. *Médecine & Droit*, 2020(161):29–34. https://doi.org/10.1016/j.meddro.2019.11.002

Eric A. Posner. 2008. Does political bias in the judiciary matter?: Implications of judicial bias studies for legal and constitutional reform. *The University of Chicago Law Review*, 75(2):853–883. https://doi.org/10.2139/ssrn.1082055

Juliano Rabelo, Mi-Young Kim, Randy Goebel, Masaharu Yoshioka, Yoshinobu Kano, and Ken Satoh. 2020. A summary of the COLIEE 2019 competition. In *New Frontiers in Artificial Intelligence*, pages 34–49. Springer International Publishing. https://doi.org/10.1007/978-3-030-58790-1_3

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.405

Joseph Raz. 1979. The rule of law and its virtue. In *The Authority of Law: Essays on Law and Morality*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198253457.001.0001

Edwina Rissland and David Skalak. 1991. Cabaret: Rule interpretation in a hybrid architecture. *International Journal of Man-Machine Studies*, 34(6):839–887. https://doi.org/10.1016/0020-7373(91)90013-W

Mark Ryan. 2020. The future of transportation: Ethical, legal, social and economic impacts of self-driving vehicles in the year 2025. *Science and Engineering Ethics*, 26(3):1185–1208. https://doi.org/10.1007/s11948-019-00130-2, PubMed: 31482471

Mari Sako, Matthias Qian, and Jacopo Attolini. 2022. Future of professional work: Evidence from legal jobs in Britain and the United States. *Journal of Professions and Organization*, 9(2):143–169. https://doi.org/10.1093/jpo/joac011

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424. https://doi.org/10.1017/S0140525X00005756

Robert Shaffer and Stephen Mayhew. 2019. Legal linking: Citation resolution and suggestion in constitutional law. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 39–44. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-2205

Shirong Shen, Guilin Qi, Zhen Li, Sheng Bi, and Lusheng Wang. 2020. Hierarchical Chinese legal event extraction via pedal attention mechanism. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 100–113. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.9

Karolina Stańczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2023. Quantifying gender bias towards politicians in cross-lingual language models. *PLOS One*, 18(11):1–24. `https://doi.org/10.1371/journal.pone.0277640`, PubMed: 38015835

Rachel Stern, Benjamin Liebman, Margaret Roberts, and Alice Wang. 2020. Automating fairness? Artificial intelligence in the Chinese courts. *Columbia Journal of Transnational Law*, 59:515.

Georgina Sturge. 2021. Court statistics for England and Wales. *House of Commons Library*.

R. E. Susskind. 2008. *The End of Lawyers?: Rethinking the Nature of Legal Services*. Oxford University Press. `https://doi.org/10.1093/oso/9780199541720.001.0001`

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.naacl-main.56`

N. N. Taleb. 2012. *Antifragile: Things that Gain from Disorder*. Penguin Books Limited.

Christopher Thomas and Antonio Ponton-Nunez. 2022. Automating judicial discretion: How algorithmic risk assessments in pretrial adjudications violate equal protection rights on the basis of race. *Law & Inequality*, 40:371.

Michael Tonry. 2014. Legal and ethical issues in the prediction of recidivism. *Federal Sentencing Reporter*, 26(3):167–176. `https://doi.org/10.1525/fsr.2014.26.3.167`

Leda Tortora, Gerben Meynen, Johannes Bijlsma, Enrico Tronci, and Stefano Ferracuti. 2020. Neuroprediction and A.I. in forensic psychiatry and criminal justice: A neurolaw perspective. *Frontiers in Psychology*, 11. `https://doi.org/10.3389/fpsyg.2020.00220`, PubMed: 32256422

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*, pages 3590–3599. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.findings-acl.314`

Kristen Vaccaro, Karrie Karahalios, Deirdre K. Mulligan, Daniel Kluttz, and Tad Hirsch. 2019. Contestability in algorithmic systems. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 523–527. Association for Computing Machinery. `https://doi.org/10.1145/3311957.3359435`

Josef Valvoda, Ryan Cotterell, and Simone Teufel. 2023. On the role of negative precedent in legal outcome prediction. *Transactions of the Association for Computational Linguistics*, 11:34–48. `https://doi.org/10.1162/tacl_a_00532`

Josef Valvoda, Tiago Pimentel, Niklas Stoehr, Ryan Cotterell, and Simone Teufel. 2021. What about the precedent: An information-theoretic analysis of common law. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2275–2288. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.naacl-main.181`

Josef Valvoda, Oliver Ray, and Ken Satoh. 2018. Using agreement statements to identify majority opinion in UKHL case law. In *Legal Knowledge and Information Systems*, pages 141–150. IOS Press.

R. Wacks. 2015. *Law: A Very Short Introduction*. Very short introductions. Oxford University Press. `https://doi.org/10.1093/actrade/9780198745624.001.0001`

Miranda Walters. 2018. London firms embrace artifical intelligence. *CBRE*.

Ziyue Wang, Baoxin Wang, Xingyi Duan, Dayong Wu, Shijin Wang, Guoping Hu, and Ting Liu. 2019. IFlyLegal: A Chinese legal system for consultation, law searching, and document analysis. In *Proceedings of the 2019 Conference*

on *Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 97–102. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-3017

Andrew Wistrich and Jeffrey John Rachlinski. 2017. Implicit bias in judicial decision making how it affects judgment and what judges can do about it. *Cornell Legal Studies Research Paper*, 17–16. https://doi.org/10.2139/ssrn.2934295

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2019. CAIL2019-SCM: A dataset of similar case matching in legal domain. *arXiv preprint arXiv:1807.02478.*

Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. 2020a. Distinguish confusing law articles for legal judgment prediction. *arXiv preprint arXiv: 2004.02557.* https://doi.org/10.18653/v1/2020.acl-main.280

Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020b. Investigating bias and fairness in facial expression recognition. In *Computer Vision – European Conference on Computer Vision 2020 Workshops*, pages 506–523. Springer International Publishing. https://doi.org/10.1007/978-3-030-65414-6_35

Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2016. Deep learning architecture for patient data de-identification in clinical records. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 32–41. The Conference on Computational Linguistics 2016 Organizing Committee.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. 2022. LEVEN: A large-scale Chinese legal event detection dataset. In *Findings of the Association for Computational Linguistics*, pages 183–201. Association for Computational Linguistics.

https://doi.org/10.18653/v1/2022.findings-acl.17

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1168

Ying Yin and Ivan Habernal. 2022. Privacy-preserving models for legal natural language processing. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 172–183. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.nllp-1.14

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1390

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020a. 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. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.466

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020b. JEC-QA: A legal-domain question answering dataset. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 9701–9708. AAAI Press. https://doi.org/10.1609/aaai.v34i05.6519

Konrad Zweigert and Hein Kötz. 1992. *Introduction to Comparative Law*. Clarendon paperbacks. Clarendon Press.