Ethical Considerations of AI in the Legal Sector

Gauri Samith (gs3280), Nicole Brye (nab2217) , Arjun Bhan(AB5666), Elie Chemouni (ec3778)

Abstract/Introduction

AI is one of the most significant technological advancements of our time, revolutionizing industries and altering the way we live. Its groundbreaking abilities to analyze and predict data far surpass those capabilities by humans. This transformative technology has found applications across various industries, from finance to healthcare to transportation. One of the most significant industries AI is affecting is the legal sector. However, the application of AI for this industry has led to some conundrums about how to effectively and ethically use this technology. Some of these concerns include, the fairness of AI models, data privacy concerns, transparency of the algorithms and biases present in training data. We will examine the ramifications of AI in the legal sector, exploring these concerns and how they will impact the field of law.

Data Curation, Privacy

The presence of the biases embedded within data have caused large ethical concerns for the development of AI systems, especially in the legal domain. The predictions that AI models make are based on the data they are trained on, and when this data contains large amounts of bias, then the model will reflect it in its results. This is particularly problematic in the legal context because our legal system is meant to safeguard individuals against societal biases and ensure there is equal justice for all. In the past, data such as police records, criminal histories and court archives have sometimes been used to discriminate against certain communities. When trained on this type of data, AI systems often perpetuate existing biases.

Although AI models excel at identifying patterns and correlations, they struggle in recognizing the complex societal inequalities that are sometimes embedded within their datasets. For example, some minority groups have historically had disproportionate rates of arrests and convictions. This trend stems partially from discriminatory practices by law enforcement that have led to systemic injustices against these groups.

The over-reliance on historical data for training AI models can reinforce biases. In a field like law, with ever changing rules and practices, it’s crucial that AI tools be trained on current laws, not outdated ones. By using historical data to train these models, it may inadvertently favor prior legal standards and policies over current ones. Additionally, it would be almost impossible for a model to keep track of how changes in the laws have influenced past decisions. However, it should be noted that most legal data comes from past laws and court decisions. This poses a challenge in training AI models, since data sets that are current and devoid of biases may be scarce.

A possible solution to this issue would be to establish a diverse multidisciplinary team of experts with technological, legal and sociology backgrounds to go through and review the data used in the model. Those from a technological background could test the data’s accuracy and relevance. Meanwhile, the experts from the legal and sociology fields could identify and remove examples from the dataset that contain social bias or legal inaccuracies. By using this multidisciplinary approach, we can attempt to create training datasets that are both unbiased and accurate.

The issues of data privacy, across various sectors and particularly within the legal system, have given rise to complex ethical dilemmas. Many companies have come under scrutiny for unethical practices, such as selling or distributing their “private” data to third parties. Some of this data may come from their customers who are often unaware of how their information is being used. This issue is very important within the legal sector, which handles sensitive data such as court details, personal records and criminal histories. This data is meant to be treated confidentially. The mishandling or leakage of this data could potentially affect legal cases giving unfair advantage to opposing parties. The leakage of client data within court hearings could compromise the interests and safety of the parties involved in the legal proceedings.

Some companies that handle consumer data have been criticized for prioritizing profit over privacy. Data brokerage firms, for example, sell ‘anonymized data’ to other organizations for profit. However, this ‘anonymization’ in some cases fails to protect data privacy, as algorithms can match individuals to their supposedly anonymous information. Furthermore, some data brokers outsource their data to other companies for ‘improved results’. This approach can lead to lax oversight and increased likelihood of data leaks potentially exposing sensitive information.

There have also been ethical concerns expressed about data labeling companies who manually tag non-numerical data (e.g., text, images, audio clips, etc.) so that AI models “understand” it. Accurate labeling of data is needed to train AI and machine learning algorithms so that they can learn how one piece of data relates to the next. There have been reports of exploitation of workers by data labeling companies. These companies exploit workers in developing countries for these tasks using their economic circumstances for profit. The integrity of labeled datasets can be compromised by using these poorly paid workers introducing biases during the data labeling process. Their subjective opinions, values, worldviews, and biases can also contribute to this issue.

In order to ensure privacy with data we must have strict regulation on how data can be shared and used for training. Personal data should be shared without consent of the individual. As data becomes more integral to our society, establishing clear guidelines and regulations for data handling is crucial. This approach ensures individual privacy rights are being followed and strengthens data security measures.

Fairness Verdicts, Accuracy Metrics, Efficiency

The integration of AI in the Legal Sector prompts a necessary examination of how these systems can be measured for fairness, how accuracy plays a crucial role in their effectiveness, and how efficiency can be balanced with ethical considerations.

Fairness in AI, particularly in legal contexts, encompasses the principle that AI systems should make decisions without biases or discrimination against any individual. In other words, an algorithm must not favor one demographic over another, which is crucial in legal decisions. However, defining fairness is not straightforward due to the subjective nature of what is considered fair in legal contexts.

First, it’s essential to translate the abstract notion of fairness into measurable indicators. A prevalent strategy involves attaining demographic parity, which dictates that the decision made by AI should not be influenced by affiliation with certain categories like race, gender or age. For example, if we consider an AI tool designed to forecast the likelihood of reoffending, this tool must ensure its predictions do not unjustly target or disadvantage any specific racial group over others. To achieve this, statistical analyses could be used to compare the distribution of outcomes among various demographic groups. This process ensures that the AI's functionality aligns with ethical standards, by verifying that its predictive outcomes are uniformly distributed across different sections of society. Specifically, when an AI application is tasked with assessing recidivism probabilities, it's critical to monitor that its forecasts do not inadvertently perpetuate racial biases, which historically have been a concern in manual legal evaluations. The method of using statistical assessments aids in identifying any disparities in how predictions are distributed across groups, highlighting potential biases.

Secondly, we can measure fairness through equality of opportunity, ensuring that individuals who are similar with respect to some relevant characteristics receive similar treatments. In legal AI applications, this might mean that all defendants with a similar criminal history and characteristics should have a similar probability of having a certain sentence, regardless of other irrelevant personal characteristics.

Another method for measuring fairness in legal AI systems involves the use of counterfactual fairness. It’s a concept that focuses on the alteration of an individual’s irrelevant characteristics without changing the outcome of the AI’s decision. This approach asks whether an AI system would have made the same decision if the individual’s irrelevant characteristic, such as race or gender, were different. For example, in the context of an AI system used for bail decisions, counterfactual fairness would evaluate whether an AI system's recommendations would remain unchanged if a defendant's protected characteristics, such as race or gender, were hypothetically altered, keeping all other variables constant. This approach tests the AI's impartiality by asking, for example, if the system would recommend bail under the same conditions if the defendant belonged to a different demographic group. If the recommendation changes due to the alteration of protected characteristics alone, the system fails the counterfactual fairness test, indicating bias. This method ensures that AI decisions in legal settings are based purely on relevant, unbiased factors, maintaining the integrity and fairness of legal judgments.

Accuracy in AI legal systems means the system's ability to make correct decisions based on the data it has been trained on. However, accuracy alone is not sufficient to deem an AI system as ethical. Indeed, an AI system could be highly accurate in predicting outcomes based on historical data, but if the data contains some biases or discriminatory practices, the AI system would perpetuate those injustices.

Therefore, measuring accuracy must be done alongside fairness metrics. For example, we can evaluate the overall accuracy of the AI system but also break down accuracy by demographic group or break down accuracy by some relevant categories. Then, the performance of the system will not benefit or harm any particular group in an important way. To deal with that, we can use confusion matrices, which allow for the examination of true positives, false positives, true negatives, and false negatives. Then, it can provide nuanced insights into where biases might exist within the system's accuracy.

Efficiency in AI legal systems can refer to the speed and resource efficiency with which these systems can process information and make decisions. But reducing the time and costs should not come at the expense of fairness and accuracy.

An efficient AI system that expedites case analysis could lead to faster resolution of legal matters which is particularly advantageous in overburdened legal systems. However, this efficiency must be carefully balanced with the need to ensure that decisions are not rushed or made with incomplete information, which could compromise fairness and accuracy.
To ensure that efficiency do not undermine ethical considerations, it needs to implement oversight mechanisms and monitoring of AI systems. For example, we can think about regular audits of AI decisions against fairness and accuracy metrics or the incorporation of feedback loops where inaccuracies or biases are corrected in subsequent iterations of the AI system.

Transparency, Accountability, Citizen Rights

It is also critical to analyze how the integration of AI in the legal sector impacts the transparency and accountability of decision making. Especially in practical applications such as predicting recidivism (which will be described in further detail below by the COMPAS case study), where AI must determine whether or not an accused defendant is likely to reoffend, it is imperative to properly explain the model decisions that will affect the livelihood of the accused and of the general population as well.

The main goals of explainability are to understand how a model works, and to help answer questions about a model's impact (Thais, 2024). For example, if an AI model predicts that a criminal may reoffend, it is crucial to determine exactly what steps in the decision-making process led to that outcome because the accused needs grounds to refute a potentially unfair decision. The consequences of a mis-classification/failing to understand the decision in this scenario, which could include imprisoning an innocent person or putting civilians in harm's way, are dire. Although perfect model explainability is currently impossible, a lack of effort on this front could allow these models to become tools for perpetuating systemic biases and for reinforcing existing inequalities.

There are four types of explainability to consider when evaluating model decisions: Local vs. Global, Example vs. Model, Individual vs. Group, and Feature vs. Mechanism. This paper delves deeper into Example vs. Model and Individual vs. Group explainability, particularly the Example and Individual aspects, as these place a heavy emphasis on explaining why individual data points/examples are assigned to specific labels. In the legal sector where model decisions have strong impact on individual livelihood, it is more sensical to place a higher emphasis on evaluating individual points rather than the overall behavior of the model. In order to provide context for the following analyses, Example vs. Model explainability refers to whether or not a specific data point or overall model behavior is being explained. Individual vs. Group explainability refers to whether or not the performance for a specific data point or category of data is being explained (Thais, 2024).

Local agnostic methods such as Local Interpretable Model agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) are ideal for evaluating the implications of AI models designed for tasks within the legal sector. With respect to the previous statements, these local agnostic methods would assist in explaining specific data points for Example explainability, and would assist in explaining the performance on a specific data point for Individual explainability. LIME requires identifying a point of interest x, and using a complex model (for example, a neural network that may not be very interpretable) to classify the remaining data points. A weight π_x is assigned to each data point concerning its proximity to x, and a series of simpler, interpretable models are trained based on the weights of the other points to generate a better explanation for how x was classified. An explainability function similar to the one below would be optimized, where f represents the complex model, g represents a simpler model, and π_x represent the weights of the points' proximities to x.

$$\varepsilon(x) = argmin_{g \in G} L(f, g, \pi) + \Omega(g)$$

In addition, SHAP is a more mathematically backed method that concerns how each individual feature affects the classification of each individual data point. The same model is trained with different combinations of features, and the performance on each individual data point after perturbing the features would look like the following:

$$SHAP_{feature}(x) = \sum_{set:feature \in set} \left[|set| \times \binom{F}{|set|}\right]^{-1} [Predict_{set}(x) - Predict_{set/feature}(x)]$$

where set represents the entire dataset, $set/feature$ represents the dataset with a specific feature removed, and x represents the point of interest. Although not perfect, these explainability methods would be recommended for helping legal professionals and civilians better understand the decision-making process, and deploying models that are *individually-focused*. This provides better backing for decisions to be refuted if needed, especially in applications described above where the consequences are serious.

With better model transparency, it becomes easier to hold the correct parties accountable for model errors. However, AI models still face major issues with accountability, especially within the legal sector; there is no designated party or third party to effectively regulate these rapidly evolving systems. As of 2023, a survey of the current policy landscape revealed that "very few algorithmic-bias focused policies have been enacted in the United States", and the frameworks that have been implemented still exhibit problematic behavior such as "distortion and discrimination" (Thais et. al). In the context of the legal sector, this implies that current frameworks responsible for holding those in charge accountable may still let issues of racial bias, socioeconomic status, and gender inequality persist. It would be highly recommended to designate an unbiased third party to monitor data collection/curation, model explainability, and the breakdown of results. This may help mitigate the issues related to rapidly evolving, under-audited AI.

Overall, civilians have the right to a fair, unbiased trial when justice is being served. The lack of transparency and accountability of the AI models that are relied upon to make legal decisions pose a huge threat to safeguarding individuals, but pushing for the use of local agnostic methods and a third party to determine accountability can mitigate the unintended consequences of model decision making.

Case Studies

Tying these concepts together, it would be appropriate to conclude with a real-life example of an application of AI in the legal sector. The aim is to get a basic understanding of what the tool is and then analyze some ethical implications regarding the tools abilities and usage. This case study focuses on the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) tool which is a risk assessment algorithm designed to evaluate the likelihood of a defendant becoming a recidivist, that is, relapsing into criminal behavior. It has been used by judges and parole officers to inform decisions on sentencing, bail and parole. Northpointe made use of multiple risk scales in order to make these predictions:

1. Pretrial Release Risk Scale: This measures whether or not the individual is likely to fail to appear or commit new felonies while on release. It makes use of current charges, prior arrest history, community ties, etc., as primary indicators.
2. General Recidivism Scale: Predicts new offenses upon release and post-assessment.
3. Violent Recidivism Scale: Predicts violent offenses post-release. This makes use of factors such as their vocational/educational problems and age.

As can be seen from the predictors of the above scores, the data used includes but is not limited to demographic information as well as social environment and community ties. Unfortunately, making use of this demographic information, especially from a historical standpoint, may inadvertently introduce biases and systemic prejudice that might have been present in law enforcement in the past. Along these lines, it becomes imperative to explore how fairness is perceived by the algorithm. As highlighted by the "(Im)possibility of Fairness", which data is appropriate to use from a historical context should be called into question. The paper details that there are two interpretations of fairness: an equality of outcomes and equality of treatment. One view presents a "what you see is what you get" approach, suggesting that the observed space is an exact reflection of the construct space (Friedler et. al). In this situation, a historical pattern of racism within a justice system would unfairly weigh against a particular racial group as highlighted. For example, a Propublica investigation claimed that COMPAS was biased against black defendants by comparing the risk assessments to actual crime records. They highlighted that black defendants were twice as likely as white defendants to be labeled high risk and did not actually re-offend and white defendants were more likely to be mislabeled at lower risk and then go on to commit further crimes. This increased the false positive rates for black defendants and the false negative rates for defendants, affecting the overall accuracy of the model (Angwin et. al).

Since COMPAS is a proprietary tool owned by a private company, the exact workings of the tool are not open to the public, making it difficult to assess fairness on a public level. Defendants and counsel are not able to challenge or understand the tool's assessments and the basis for the specific risk scores. This leaves little room for accountability in a domain like the justice system, where it is paramount to be able to appeal against decisions made. This may also have an impact on the actual sentencing process. Blind faith in a black box tool such as COMPAS may encourage an over-reliance on AI tools which diminishes the role of judicial discretion that comes with a jury analysis based on human judgment and the unique circumstances of a case. It could normalize the previously mentioned systemic biases against particular groups in society and invoke feelings of distrust in the justice system among these communities. It is necessary to address and be aware of the structural biases and possible weaknesses of the tool in order to make informed decisions. The scores generated by the model should not influence the actual beliefs of a jury and should merely serve as a tool of confirmation rather than the final decision maker in the process.

References

1. Thais, S., Shumway, H., Saragih, A. I. Algorithmic bias: Looking beyond data bias to ensure algorithmic accountability and equity. MIT Science Policy Review 4, 59–66 (2023).
2. Thais, S. (2024). *Lecture 8: Current Applications of AI* [PowerPoint slides]. Columbia University Data Science Capstone and Ethics
3. Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *arXiv:1703.09207v1*, (2017).
4. Kusner, M., Lofts, J., Russell, C., and Silva, R. Counterfactual fairness. arXiv:1703.06856, (2017).
5. Thais, S. An Introduction to AI Ethics and Responsible Data Science. (2024).
6. Thais, S. Algorithmic Bias and Quantitative Fairness. (2024).
7. Thais, S. Explainable/Interpretable AI. (2024).
8. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

9. Starr, S. B. (2014). Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. Stanford Law Review, 66(4), 803-872. Retrieved from http://www.stanfordlawreview.org/wp-content/uploads/sites/3/2014/05/66_Stan_L_Rev_803_Starr.pdf

10. Barabas, C., Dinakar, K., Ito, J., Virza, M., & Zittrain, J. (2018). Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. Conference on Fairness, Accountability, and Transparency (FAT) 2018. Retrieved from https://arxiv.org/abs/1711.08230

11. Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016, September 23). On the (im)possibility of fairness. arXiv preprint arXiv:1609.07236. Retrieved from https://arxiv.org/abs/1609.07236