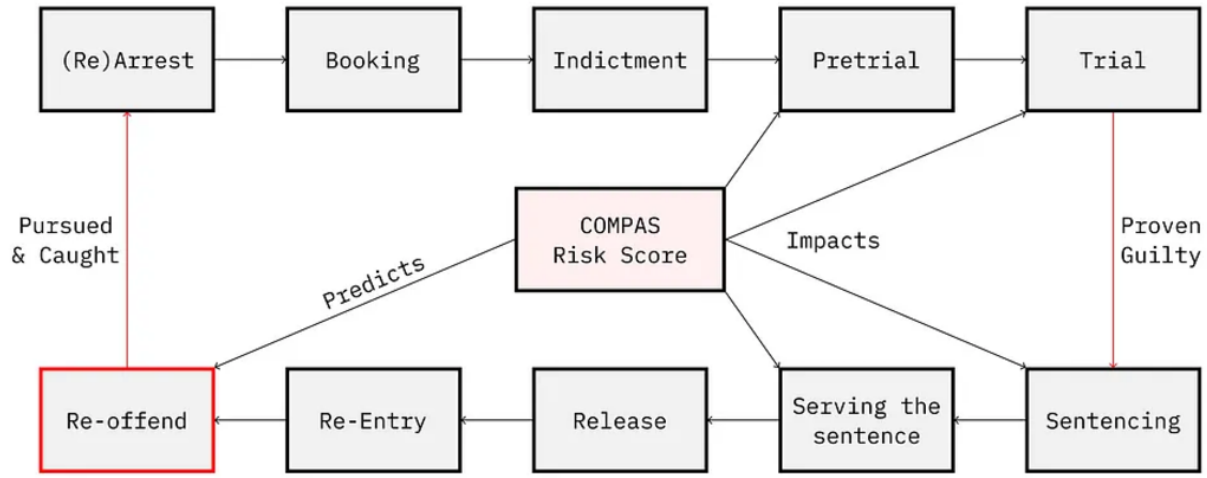
AI for Recidivism and Judicial outcomes

Department of master's in data science, Monmouth University, West Long Branch, USA.

Sushika Reddy Gade

Introduction:

The use of AI models are used in predicting recidivism, i.e., the likelihood of a defendant committing another crime. These AI systems are used in judicial decision-making processes, such as sentencing and bail decisions, where they aim to provide objective assessments.



A simplified overview of steps in the criminal justice process[10] and how COMPAS scores relate to them

Algorithm named COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), was developed by Northpointe (now Equivant) to predict the likelihood of a criminal defendant to commit another crime. This algorithm, developed for courts and correction agencies for determining the question of bail, sentencing, and parole, considers scores of various variables, including criminal history, substance use, living situation, and socioeconomic background, and returns a risk score ranging from 1 to 10.

In 2016, when ProPublica discovered that the algorithm discriminated against blacks, flagging them as potential recidivists at nearly twice the rate as whites, the company denied the results. This in turn brought a great deal of controversy and criticism to COMPAS. This case raised highly important issues of algorithmic bias, openness of criminal justice tools, and wider ethical implications of using AI systems for high-stakes decision-making.

1. Why is this an ethical problem?

1.1 Bias and Fairness:

These algorithms have been shown to misclassify risk levels, with higher risk scores assigned disproportionately to Black defendants compared to White defendants, despite similar profiles. This reflects systemic racial bias in AI decision-making. Socio-economic and age-specific biases also influence predictions, potentially perpetuating existing inequalities.

1.2 Accountability and Explainability:

Algorithms like COMPAS lack transparency, as their decision-making processes are often proprietary and not open to scrutiny. This reduces trust and makes it difficult to hold the systems accountable for unfair outcomes.

1.3 Impact on Judicial Decisions:

Judges may rely heavily on AI predictions without fully understanding the limitations or biases in the model. This can lead to unjust outcomes, such as harsher sentences or unfair denial of parole for certain groups.

1.4 Non-Maleficence:

AI systems must "do no harm," but when biased, they harm individuals by reinforcing societal discrimination and affecting their life trajectories.

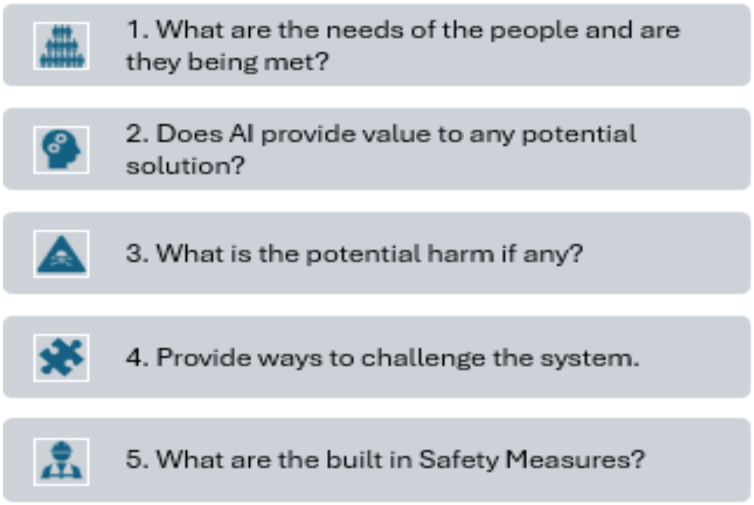
1.5 Human Rights Concerns:

The decisions influenced by biased AI can undermine fundamental rights, such as equality and fair treatment under the law.

1.6 Ethical Importance:

Addressing these issues is crucial to ensure that AI tools in judicial systems uphold fairness, justice, and transparency, mitigating potential biases and harm and supporting equitable legal outcomes. This requires designing AI models aligned with ethical principles and integrating human oversight into critical decision-making processes.

Use of Human centered Design helps in tackling most of the above-mentioned problems and the model aligns with societal as well as judicial ethics.



Ensuring Human intervention and safety measures in the decision-making process and also providing ways to challenge the system and it's outcomes.

2. Does the Current Business Case Demonstrate Accountability?

The current business case falls short of demonstrating full accountability.

Why not?

2.1 Lack of Transparent Model Documentation:
While the case highlights goals like fairness, bias mitigation, and accountability, it doesn't include detailed documentation, such as a model card. This type of documentation is essential for explaining how the AI makes decisions, what its limitations are, and what biases it might have.

2.2 Bias Issues:
The case acknowledges racial bias in COMPAS but doesn't provide enough detail about steps for ongoing bias monitoring or independent audits. These are critical to ensuring that biases are consistently addressed over time.

2.3 Limited Explainability:
Although transparency is mentioned as a goal, there's no clear plan for how to make the AI's decisions understandable. For instance, it's unclear what

methods or tools, like SHAP or LIME, will be used to explain the model's outputs.

2.4 No Clear Oversight Mechanism:
Accountability requires a well-defined governance structure. The current case lacks a framework for assigning responsibilities to stakeholders who can intervene when AI produces questionable decisions.

Recommendations To build a Transparent AI Judicial System

Develop Comprehensive Model Documentation:
Create a model card that explains the AI's purpose, the data it uses, its performance, known limitations, and potential biases.

Conduct Regular Bias Audits:
Schedule periodic reviews to identify and address disparities in the model's outcomes, particularly related to race, socio-economic status, or gender.

Implement Explainability Tools:
Use tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) to make the AI's decisions clear and understandable for stakeholders, including judges.

Establish a Governance Framework:
Define roles for data scientists, legal experts, and ethicists to oversee the AI system. This team should have the authority to address concerns and ensure the model is used responsibly.

Incorporate Human Oversight:
Require human review for high-stakes decisions. Judges and decision-makers should critically evaluate the AI's recommendations rather than relying on them blindly.

By implementing these steps, the business case can better demonstrate accountability and ensure the AI system is used in a fair, transparent, and responsible way.

Algorithmic Impact Assessment (AIA): AIA stands for Algorithmic Impact Assessment. It is a full study of the impacts of the model on society. As this application can severely cause harm to a patient, possibly kill them, it is important in understanding the impacts prior to launch. To understand the full impact this may have on society, we will be using Canada's model for AIA. This involves developing a questionnaire form for all users, judges, recidivists along with a scoring system of Impact Levels I-IV to determine the overall impact of the model.

s.no	Mock Algorithmic Impact Assessment:
1.	Is there a need for a tool to predict recidivism?
2.	Have you considered recidivism an ethical issue?
3.	What are your thoughts on technology mitigating bias within judicial systems and decision making?
4.	Would the platform influence your judgement plan for the person?
5.	Has current technology for predicting recidivism proven beneficial to judicial practices?
6.	What concerns do you have regarding privacy of your personal information?
7.	Do you think that this application will serve its purpose well?
8.	Are there any parts of the AI that you do not trust to assist in predicting recidivism?

With this preparatory review, we will understand everyone's needs even before the development begins. The process will allow all potential users to understand the roles and benefits that our recidivism prediction tool will provide.

3. Does the Current Business Case Demonstrate Transparency?

The current system lacks sufficient transparency.

3.1 Here's Why:

Black Box Nature of COMPAS:

The COMPAS algorithm, used to predict recidivism, operates like a black

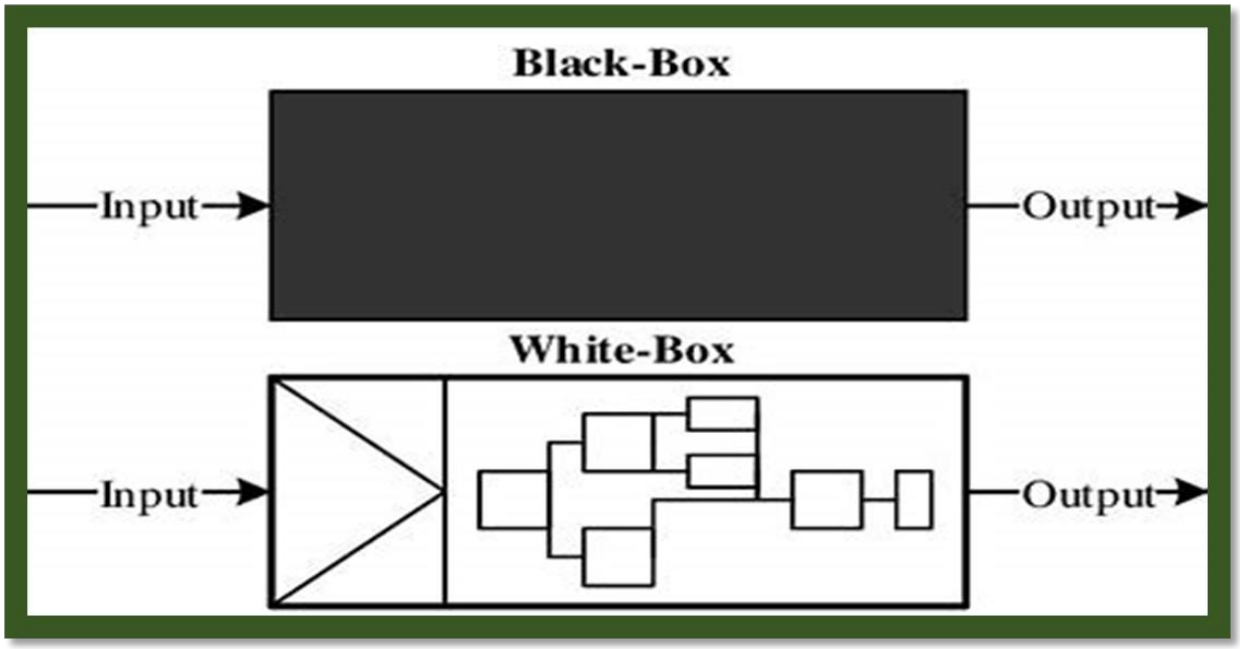
box. Its design is proprietary, so we don't know how it works or what factors influence its predictions. This lack of openness makes it impossible to understand how decisions are made.

Hard to Interpret:
The way the algorithm makes decisions isn't easy for non-technical people—like judges or defendants—to understand. This makes it harder to trust the system because users can't verify whether the predictions are accurate or fair.

No Tools for Explanation:
There's no mention of using tools like SHAP (SHAPley Additive explanations) or LIME (Local Interpretable Model-agnostic Explanations) to clarify how the model produces its results.

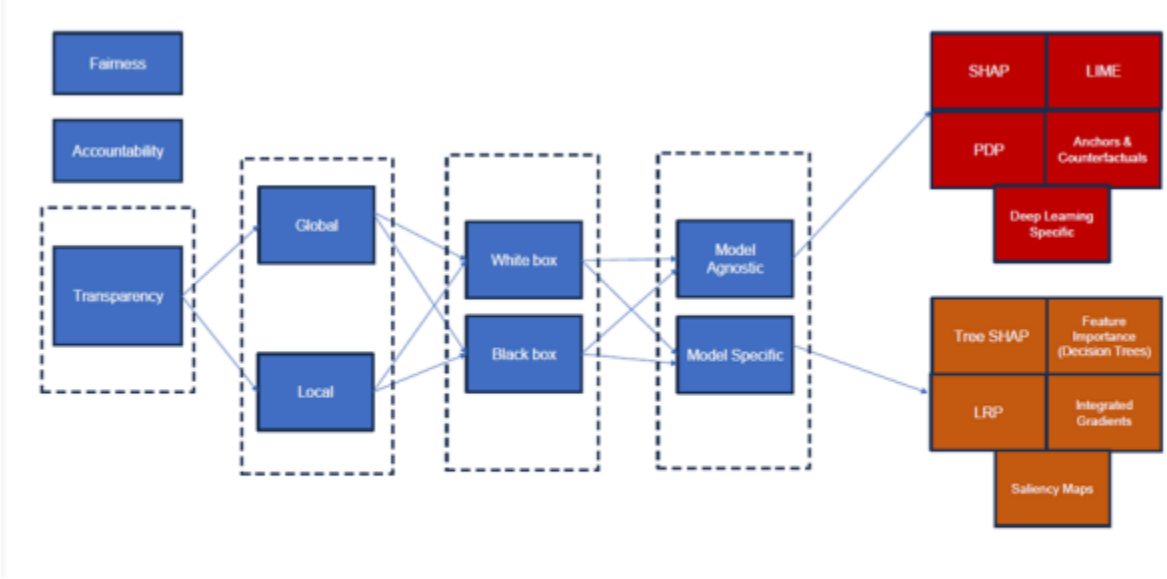
Unclear Data Handling:
It's not clear how the data used in the system was chosen, processed, or weighted. This raises concerns about whether biases in the training data have been addressed properly.

Black Box vs. White Box Algorithms



- **Black Box Algorithms:**
  These systems, like COMPAS, prioritize complex computations (e.g., neural networks) but sacrifice interpretability, making them difficult to scrutinize.
- **White Box Algorithms:**
  Transparent models (e.g., decision trees, linear models) allow users to trace how inputs lead to outputs, which is essential in high-stakes contexts like judicial decision-making.

3.2 For Improved Transparency



"Decision matrix for model interpretation techniques, showcasing tools like SHAP and LIME and their application in achieving transparency, fairness, and accountability in AI models."

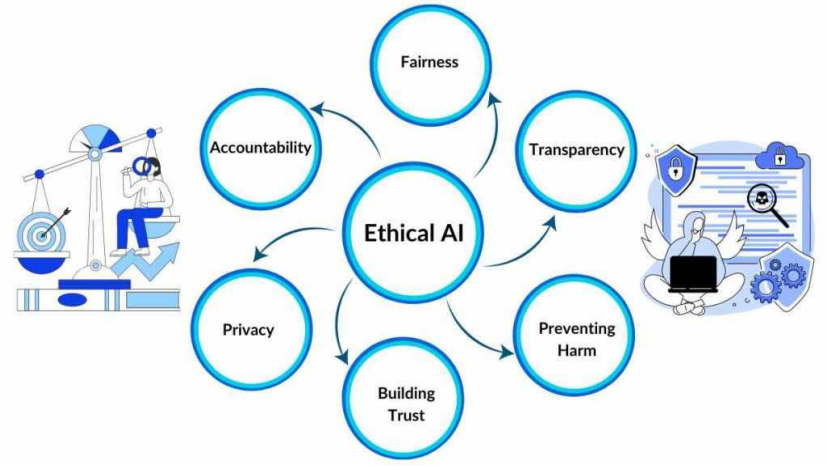
1. **Use Interpretable Models:**
   Whenever possible, choose algorithms like decision trees, rule-based systems, or linear models or for a matter of fact we can still use XG-Boost in comparison to other neural networks. These models make it easy to see how decisions are made and why specific outcomes are reached.
2. **Add Explanation Tools:**
   If using black box models is necessary, integrate tools that clarify how predictions are generated:

- **SHAP (Shapley Additive Explanations):** Highlights which features contribute most to a prediction.
- **LIME (Local Interpretable Model-Agnostic Explanations):** Simplifies individual predictions so they are easier to understand.

3. **Open Up the Model Design:**
   Share clear documentation about how the algorithm works. This should include details like the input features, how they are weighted, and how these factors impact predictions.
4. **Use Transparent Datasets:**
   When possible, make the training data available for review and document every preprocessing step. This ensures biases are identified and addressed.
5. **Introduce Model Cards:**
   Create a "model card" that explains the algorithm's purpose, the data used for training, its performance metrics, limitations, and any ethical safeguards in place.
6. **Conduct Regular Audits:**
   Schedule regular reviews to ensure the algorithm stays unbiased and its decision-making remains clear over time.

These steps can make the system more transparent, fostering trust and accountability among all stakeholders

4. Privacy Regulations and Citizen Rights



4.1 Following Federal and State Laws:

- **At the Federal Level:**
    - Comply with the Privacy Act of 1974 to ensure ethical practices in collecting and using personal information.
    - Follow AI-specific guidelines from agencies like the Federal Trade Commission (FTC) to prevent deceptive or unfair practices.
- **At the State Level:**
    - Adhere to the California Consumer Privacy Act (CCPA) and its updated version, the California Privacy Rights Act (CPRA). These laws set a strong standard for data privacy and often serve as a model for other states.

4.2 Citizen Opt-In and Opt-Out Options:

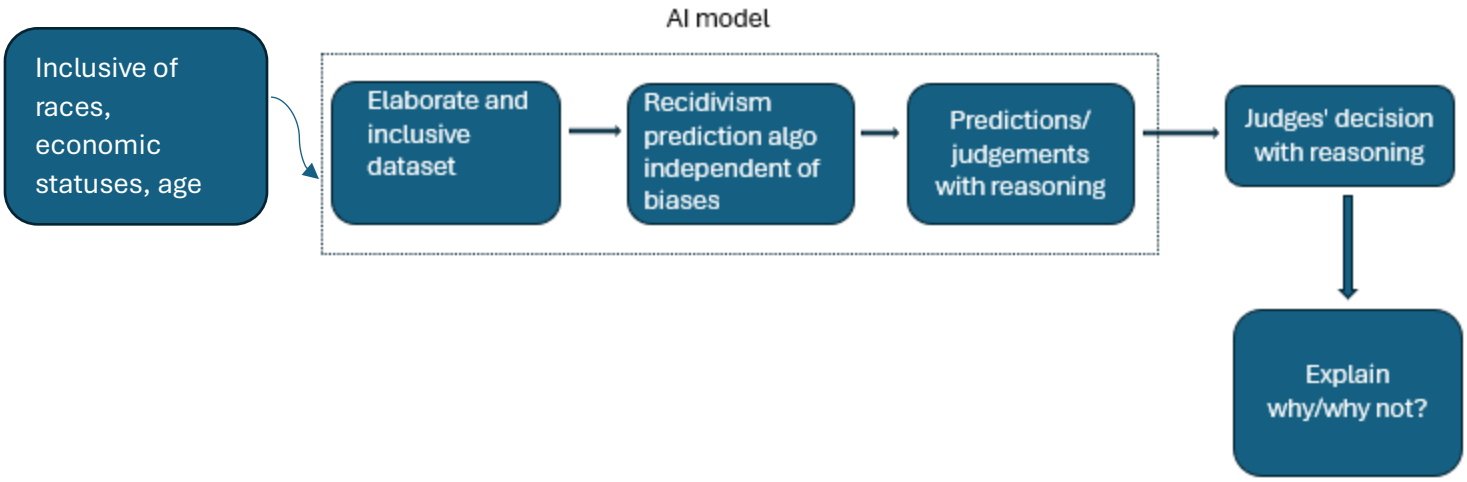
- **Opt-In Policy:**
  Ensure users actively agree to share their data, especially sensitive details like criminal records or socio-economic information.
- **Opt-Out Policy:**
  Offer citizens simple, accessible ways to withdraw consent or request the deletion of their data. This could be through an easy-to-use online portal or by contacting judicial authorities directly.

4.3 Data Minimization and Anonymization:

- Collect only the information necessary for recidivism predictions, avoiding unnecessary data.
- Use anonymization techniques to protect individuals' identities, ensuring personal identifiers are either removed or encrypted.

5. AI Fairness and BIAS

5.1 High level Diagram of the Recommended platform.



When working with judicial data, measurement bias is an inevitable challenge. Our AI system may not be able to fully eliminate biases related to race, ethnicity, or other factors. While removing demographic data like race and gender could reduce bias, it's important to recognize that this information provides valuable context for understanding the systemic factors that shape judicial outcomes.

For instance, socioeconomic conditions in certain communities can influence recidivism rates, and ignoring these factors might oversimplify the decision-making process. Instead of removing demographic data entirely, we aim to balance fairness and context, ensuring that the system can make informed, nuanced decisions.

To ensure a fair evaluation and prevent bias, our judicial AI model will first be implemented within the specific court system it was designed for. U.S. privacy and confidentiality laws prevent judicial data from being shared freely between courts or jurisdictions. If the model proves effective and beneficial in this initial setting, it can serve as a framework for other courts. Each court can then input its own data to adapt the model for localized use, ensuring it complies with legal and ethical standards.

Aggregation bias happens when different groups in the dataset are mistakenly combined, which can cause the model to perform well only for the majority group or certain subgroups. To prevent this, we will keep demographic groups separated during the model training phase. This approach is important because certain patterns, like recidivism rates or sentencing outcomes, may affect specific ethnic or socioeconomic groups more than others. By separating these groups during training, we can ensure that the model delivers fair and accurate results for everyone in the judicial system, rather than favoring one group over another

The final type of bias is deployment bias, which arises when human judgment influences the output of the model. In the context of the judicial system, even with AI support, moral and legal responsibility ultimately lies with the judge or decision-maker. Their final interpretation of all the defendant's information, combined with the output of the AI model, will determine how much bias persists in this sensitive process.

5.2 Fairness in Judicial AI

Fairness is a vital aspect of any AI application, especially in the judicial system. While it’s challenging to meet every fairness criterion—like demographic parity, equal opportunity, equal accuracy, and group unawareness, the system strives to uphold some of these principles.

For example, the AI platform promotes **equal opportunity** by standardizing sentencing recommendations across cases. This ensures that individuals, regardless of their background, are treated fairly, aligning with the core judicial ideal that everyone deserves a fair trial. Additionally, ongoing efforts to introduce standardized sentencing guidelines further support fairness and help reduce disparities.

Another principle the system addresses is **demographic party.** By using diverse and representative training data, the system aims to ensure that all racial, ethnic, and socioeconomic groups are proportionally included. This requires a thoughtful approach during data collection, especially when dealing with historical biases or underrepresented groups.

While no model or dataset can be entirely free of bias, the platform is designed to minimize bias at every stage—data collection, model training, and deployment. Through regular audits, stakeholder input, and updates, the system continually evolves to deliver outcomes that are as fair and equitable as possible for judicial use.

6. Explainability

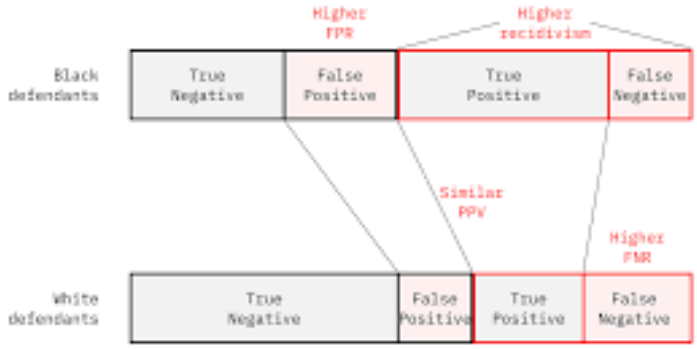
It's important to also highlight the concept of explainability, which is crucial for any AI application. There's a key difference between white box and black box models when it comes to how easily their decisions can be explained. White box models are more transparent, allowing individuals to easily trace and understand the decision-making process. On the other hand, black box models, like the one used in our judicial AI system, are harder to explain because their decision-making processes are not as transparent and can't be easily broken down by hand.

Since our system relies on a black box model, we won't be able to fully explain how it arrived at specific outcomes in a way that everyone can easily understand. This lack of transparency might raise concerns, especially for legal professionals or defendants who need to justify or challenge a decision.

To help address this, we focus on two types of explanations:

- **Local Explanations**: These explain the reasoning behind individual predictions, such as a specific sentence or parole recommendation for a defendant.

- **Global Explanations**: These focus on the overall behavior of the model, helping to explain how it operates across all cases and decisions. Providing both local and global explanations is essential to ensure the model's predictions are fair, valid, and aligned with legal and ethical standards. These explanations are key to ensuring the system is accountable and fair, ultimately supporting the judicial process.



To ensure flexibility in explaining the decisions made by our judicial AI model, use of SHapley Additive exPlanations (SHAP). SHAP is a model-agnostic method, meaning it can explain the outcomes of various machine learning models used in our system. This flexibility is a key advantage, as it enables us to apply it across different stages of judicial decision-making.

SHAP is based on game theory, treating each data point (like an individual case or defendant's profile) as a "game." This approach works well for providing local interpretations, helping us explain the reasoning behind specific outcomes, such as sentencing recommendations or parole decisions.

7. Final Recommendation

Ethical Principle	Current Challenges	Proposed Solutions
Fairness	Risk of racially, socioeconomically, or gender-biased forecasts in court decisions.	Ensure that predictions are fair and unbiased for all parties, irrespective of race, socio-economic status, or sex, by training on diverse and balanced datasets and conducting regular fairness audits
Accountability	Lack of transparency and clarity around the AI's decision-making process.	Provide a clear and elaborate model card that showcases how the model was built, including the data used, performance metrics, and the steps taken to ensure

		accountability in the model's outcomes
Bias	Potential bias in the model due to imbalances or discrepancies in the training data	Train the model on comprehensive datasets to address and correct any discrepancies, ensuring the AI makes unbiased decisions that reflect all groups fairly. Regularly update the model to minimize bias in outcomes.
Transparency	Opaque decision-making processes due to the black-box nature of the model	Provide clear and detailed explanations for every decision, including why and how a particular decision (e.g., recidivism prediction or sentencing) was made. Use interpretability methods like SHAP and LIME to make predictions understandable and explainable
Data Privacy and Security	Risks related to unauthorized access or misuse of sensitive judicial data	Adhere strictly to privacy and security laws, ensuring that only necessary data is collected and used. Implement encryption and anonymization

		techniques to protect sensitive information and comply with data privacy regulations like GDPR and HIPAA

References

1. https://www.google.com/imgres?q=force%20plot%20for%20compas%20algorithm&imgurl=https%3A%2F%2Fmiro.medium.com%2Fv2%2Fresize%3Afit%3A2000%2F1*A1oX2ICA2A7syrWxneRcWQ.png&imgrefurl=https%3A%2F%2Fmedium.com%2F%40lamdaa%2Fcompas-unfair-algorithm-812702ed6a6a&docid=9fY5FzLW3FcjYM&tbnid=7ZIbUQA_ct-FIM&vet=12ahUKEwi3koPAupmKAxVcFFkFHX0lIQQQM3oECE0QAA..i&w=2000&h=760&hcb=2&ved=2ahUKEwi3koPAupmKAxVcFFkFHX0lIQQQM3oECE0QAA
2. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algor
3. https://pbiecek.github.io/xai_stories/story-heloc-credits.html
4. https://pbiecek.github.io/xai_stories/story-compas.html#model-specific
5. https://www.google.com/url?sa=i&url=https%3A%2F%2Fai.plainenglish.io%2Fblack-box-models-vs-white-box-models-68e919170b94&psig=AOvVaw0iQhbef0lstlgWt93iGnm0&ust=1733786700793000&source=images&cd=vfe&opi=89978449&ved=0CBUQjRxqFwoTCKiM__iomYoDFQAAAAAdAAAAABAt
6. https://www.google.com/imgres?q=force%20plot%20for%20compas%20algorithm&imgurl=https%3A%2F%2Fmiro.medium.com%2Fv2%2Fresize%3Afit%3A2000%2F1*beTnYXQrHzubR3v6L04k5w.png&imgrefurl=https%3A%2F%2Fmedium.com%2F%40lamdaa%2Fcompas-unfair-algorithm-812702ed6a6a&docid=9fY5FzLW3FcjYM&tbnid=FTqhllUxbySPuM&vet=12ahUKEwi3koPAupmKAxVcFFkFHX0lIQQQM3oECF8QAA..i&w=2000&h=820&hcb=2&ved=2ahUKEwi3koPAupmKAxVcFFkFHX0lIQQQM3oECF8QAA
7. SHAP Values for Multi-Output Regression Models — SHAP latest documentation
8. Professor Arup Das notes.