IMPLICATIONS OF BLACK BOX DILEMMA IN THE INDIAN LEGAL SYSTEM

Amandeep Singh[*] Janees Rafiq[*]

ABSTRACT

The integration of AI in judicial proceedings through predictive justice tools is reshaping the decision-making process with risk assessment, recidivism forecasting, and sentencing prediction for the efficiency and consistency of the judicial system. Predictive justice refers to using analysis of a large amount of data by means of AI-enabled technologies for predicting outcomes of legal disputes. However, the growing role of AI in assisting decisions in courts is also raising critical concerns about transparency, accountability, and fairness due to the black box nature of these technologies. The black box dilemma refers to our inability to understand how deep learning systems arrive at their decisions. The opacity in these technologies raises serious ethical and legal concerns, such as biases, discrimination, and accountability gaps in various AI-assisted court decisions, which make the judicial system vulnerable to systematic biases perpetuated by inbuilt AI algorithms. These algorithms are yet to be governed by suitable legislation, which is where the gap lies. This paper showcases all the challenges and problems associated with the black box nature of technologies in the AI-assisted legal system. This paper advocates for technical solutions like explainable AI (XAI) to ensure transparency, algorithm auditing standards to detect bias, and liability frameworks to establish accountability. The paper further evaluates various global frameworks to suggest the incorporation of best practices from around the world into our legal system. The paper further provides policy-driven solutions to align AI integration with the legal system and focuses on technological innovation with ethical governance from a broader perspective.

Keywords: Artificial Intelligence, Legal System, Blackbox Dilemma, Integration, Accountability.

[*]LLB, FIRST YEAR, MODEL INSTITUTE OF ENGINEERING AND TECHNOLOGY, JAMMU.
[*]MODEL INSTITUTE OF ENGINEERING AND TECHNOLOGY, JAMMU.

INTRODUCTION

The Black Box problem is a severe issue for transparency, fairness, and accountability in an AI-powered legal system. Through techniques such as Explainable AI (XAI), algorithmic auditing, and strict regulations, transparency can be enhanced and bias reduced in AI systems. A clear set of regulations for AI regulation will make sure that AI application in the legal system is supported by strong ethical and legal foundations.1 Courts around various legal systems are employing AI tools to assist judges in predicting sentences and recidivism, with the hope of decreasing case backlogs and promoting fairness in legal decisions.2 However, the technology poses new ethical and legal dilemmas. The Black Box issue comes up when AI algorithms act in a way that is hard to understand, and it becomes very difficult to comprehend how the decisions are being made, which is a gigantic problem for transparency, accountability, and fairness in judicial proceedings.3

The potential for bias in AI-created legal decisions, as well as the difficulty of challenging or reviewing these computerized decisions, is a serious threat to the principles of justice and due process.4 This study provides a critical analysis of the implications involved with the Black Box problem in the context of AI-supported legal decision-making, with an emphasis on the Indian legal framework. It analyzes how the secrecy associated with AI systems can be detrimental to the integrity of the judiciary and public confidence.5 Through a comprehensive review of global AI governance frameworks, including the EU AI Act, GDPR, and the Algorithmic Accountability Act (USA), this study aims to determine best practices that facilitate AI-informed legal findings being transparent and equitable.6 It also promotes the adoption of practices like Explainable AI (XAI), algorithmic audits, and regulatory oversight to achieve a balance between technological efficiency and essential legal rights.

---

[1] Yavar Bathaee, 'The Artificial Intelligence Black Box and the Failure of Intent and Causation' (2018) 31 Harvard Journal of Law & Technology 889.
[2] Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (2015) (Harvard University Press 2015).
[3] Julia Angwin and others, 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks' (ProPublica, 23 May 2016) https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing accessed 5 February 2025
[4] Ibid
[5] Stanley Greenstein, 'Preserving the Rule of Law in the Era of Artificial Intelligence' (2021) International Journal of Law and Information Technology 137.
[6] European Parliament, 'The EU Artificial Intelligence Act: A Risk-Based Framework for AI Regulation' (2024) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 accessed 5 February 2025.

As artificial intelligence continues to shape the future of the judicial process, the following question remains valid: Is it possible to trust a system that resists full understanding? This research seeks to answer this question by offering policy-informed recommendations intended to ensure that AI augments, not undermines, the fairness and accountability that are inherent in judicial adjudication.

UNDERSTANDING THE AI AND BLACK BOX DILEMMA

Various official definitions of artificial intelligence systems exist, originating from independent organisations and national legislators. To provide a foundational reference, I will draw upon the definition given in the European Union Artificial Intelligence ACT 2024 (EU AI Act 2024), *" AI system means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."*[7]

The key elements in this definition are 'infers' and 'autonomy', which make an AI system different from traditional software, where the outcome is pre-defined (if a then b) by a strict algorithm. For example, a virtual assistant like Siri or Alexa infers user intent and improves over time, while a basic voice command system only responds to pre-set phrases. The definition ensures the Act remains relevant over time as the AI evolves by adopting a broad and technology-neutral approach.[8]

There is a difference between the AI that learns from data and the AI that simply evaluates possible outcomes. Consider the following hypothetical computer program designed to filter a resume for a job profile. The program is tasked with reviewing the applicant's resume and making decisions based on the applicant's qualifications, experience, and specific skills. The traditional AI or strict algorithm will follow pre-defined rules to shortlist candidates. For example:

(i) If an applicant has a minimum of 3 years of experience, they pass.

[7] European Union Artificial Intelligence Act (EU AI Act) 2023, Art. 3(1)
[8] The EU Artificial Intelligence Act: our 16 key takeaways. Retrieved from https://www.stibbe.com/publications-and-insights/the-eu-artificial-intelligence-act-our-16-key-takeaways (visited on 6th February, 2025)

(ii) If an applicant has a Master's degree, they pass.

(iii) If an Applicant has specific keywords like "Python" or "Project Management" in their resume, they pass.

Whereas, the Modern AI or learning-based AI system is given the same data given to the previous system, along with historical hiring data and the performance in past experiences. Instead of following predefined rules, this system learns patterns from the historical data to make better decisions on its own. For example, this system might prefer a candidate who has performed well in a role relevant to a Master's degree, even if they don't possess the degree, or it might favour consistent career growth and leadership qualities, even if the applicant hasn't explicitly mentioned them.

This paper is concerned with the Modern AI or learning-based AI system that learns from data and solves problems dynamically. These systems use machine-learning algorithms, as the one described in the above example, to arrive at a dynamic solution to a problem.[9]

HOW DO MACHINE LEARNING ALGORITHMS WORK?

Machine learning is a branch of Artificial Intelligence (AI) that enables systems to learn from data and make predictions independently, without direct programming or pre-established rules.[10] To understand how these machine learning algorithms work, let us take an example where an AI system is used to predict whether a defendant is likely to re-offend or not. This is a common application being used in courts to assist judges in making informed decisions about bail, sentencing, or parole.

Step 1: Data Collection: The initial step is to gather relevant data. For this example, the data set could include the accused's background, such as age, criminal record, education level, and previous convictions for recidivism. For instance:

(a) Defendant A: Age 25, 2 previous convictions, working, high school diploma, reoffended within 2 years.

[9] Yavar Bathaee, "The Artificial Intelligence Black Box and Failure of Intent and Causation", (2018) Harvard Journal of Law & Technology, Volume 31, No. 2 Spring

[10] GeeksforGeeks, "Machine Learning Tutorial" (2025) Retrieved from https://www.geeksforgeeks.org/machine-learning

(b) Defendant B: Age 40, no previous convictions, not working, college graduate, did not reoffend.

This information is utilized to train the machine learning algorithm to identify patterns common with recidivism.

Step 2: Data Pre-processing: Raw data has to be cleaned and converted into a usable format since the data is usually incomplete or inconsistent. For instance, missing values or categorical data, such as education level or criminal record, need to be converted into numerical values. This prepares the data for analysis but also introduces an additional layer of complexity that renders the final decision traceless.

Step 3: Model Training: The algorithm is then trained on the pre-processed data to identify patterns and relationships. For example, the model can learn that those accused with prior convictions and low education levels are more likely to re-offend. However, the training process is done through millions of calculations by interpreting the large number of such past cases, which makes the model's decision-making process difficult to interpret.

Step 4: Model Evaluation: The trained model is then tested on another dataset to test its performance. Different metrics are used to measure the accuracy of the model for the new dataset. The model may have high accuracy in predicting recidivism, but the metric does not tell us how the model arrived at its prediction.

Step 5: Deployment: Once the model has high accuracy and reliability then it can be deployed in real-world applications. For example, in real AI systems like COMPAS and HART, a judge may use the model's prediction to decide whether the accused should be granted bail or not.

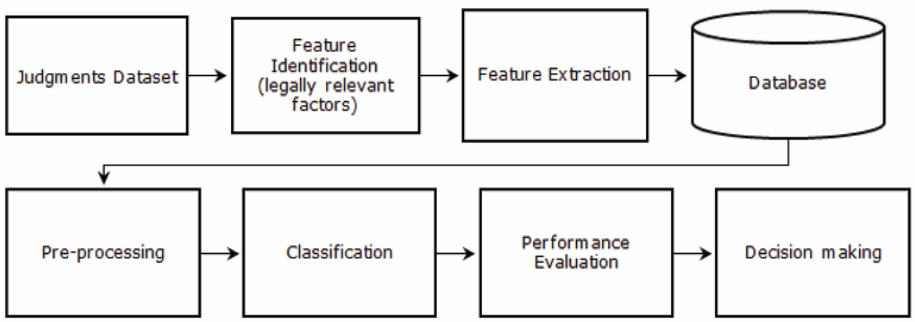

Figure 1: Research Methodology for developing predictive models. (Sheikh et. al.2020)

However, the lack of transparency in AI systems due to the complexity of algorithm structure and the ability of machine learning algorithms to make decisions on their own by internalizing data in ways that are not easily audited or understood by humans makes AI systems a black box to humans.

WHAT IS THE BLACK BOX DILEMMA?

Generally, the Blackbox Dilemma can be defined as our inability to fully understand an AI's decision-making process and the inability to predict its outputs or decisions. Imagine a magic trick in which you'll see the start and finish, but the method remains hidden. This is the essence of the Black Box problem. The lack of transparency in AI systems determines how much humans can interpret or audit the decision-making process of these systems.

As discussed in the previous section, the AI system learns from data. It identifies patterns, correlations, and relationships within vast datasets. Based on this learning, it makes predictions or decisions. An AI system adjusts its parameters guided by algorithms and feedback from the data it processes. These systems using deep learning models operate with a level of complexity that makes it difficult to decipher their decision-making logic.

This opacity poses challenges like how we can trust a system we don't understand and how we can hold it accountable for its actions. These are the critical questions as AI becomes increasingly integrated into our legal system. The Black Box problem is not just a technical challenge but a societal one with implications for fairness, accountability and safety.

In an analysis of Northpointe's tool, called "COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions), it was found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk."[11]

The black box problem presents significant risks, primarily by undermining the trust in AI systems. For instance, if a self-driving car suddenly brakes without any reason, it would be very difficult to trust the car's ability to navigate safely. Trust is crucial for the adoption and

[11] Jeff Larson and others," How we Analyzed the COMPAS Recidivism Algorithm". Retrieved from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

acceptance of any technology, especially those that have the potential to impact our lives so profoundly.

The UK's Harm and Assessment Risk Tool (HART) was intended to forecast criminal risk levels, yet research has found that it could misclassify suspects on racial or economic grounds.[12]

India's Supreme Court introduced SUPACE, an artificial intelligence-based legal research system, to enable judges to access legal files, identify relevant previous cases, and deliver judgments more quickly.[13] SUPACE applies machine learning to process large volumes of legal data to enable judges to make improved decisions. There are apprehensions about a lack of transparency that may result in biased recommendations, failure to identify significant case laws, and excessive dependence on AI for legal arguments.[14]

SUVAS is a computer translation tool based on AI that helps in translating legal documents and court judgments in many Indian languages. As India is a country with many different languages, this tool aims to make legal information easily understandable for people who don't speak English. Being based on NLP algorithms, AI translations sometimes translate legal lingo inaccurately, change the meaning of court judgments, or the translation fails.[15] Without human verification of the translated legal documents, there could be ambiguity and possible wrong legal interpretations.[16]

CMS is an artificial intelligence platform used in courts for case filing automation, pending case tracking, and hearing scheduling.[17] It can potentially minimise delays in case processing and enhance Indian court efficiency. AI-based case prioritisation and scheduling automation could be unexplainable in decision-making and cause unintended bias in case management.[18]

[12] Christopher Markou and Huw Roberts, 'AI and Criminal Justice: The Use of HART in the UK' (2023) 45(2) Cambridge Law Review 178
[13] Supreme Court of India, 'SUPACE: AI-Based Legal Research Tool' (2021) https://main.sci.gov.in/ accessed 6 February 2025.
[14] Rishika Jain, 'Artificial Intelligence in the Indian Judiciary: A Critical Analysis of SUPACE' (2022) 7(1) Indian Journal of Law and Technology 82.
[15] MeitY, 'SUVAS - AI-Based Language Translation Tool for Courts' (2020) https://www.meity.gov.in/ accessed 18 February 2025.
[16] Ankit Srivastava, 'AI and NLP in Indian Legal Translation: Challenges and Opportunities' (2023) 9(2) International Journal of Linguistics and AI 52.
[17] Ministry of Law and Justice, 'CMS Implementation in Indian Courts' (2022) https://lawmin.gov.in/ accessed 18 February 2025
[18] Arvind Datar, 'Algorithmic Bias in Judicial AI: Concerns with CMS' (2023) 12(3) NUJS Law Review 128.

When AI systems give priority to some types of cases without explainability, they can compromise the principle of equal access to justice.[19]

NATGRID is an AI-driven surveillance system that collects data from various security organisations to monitor potential security threats.[20] It applies AI software to monitor vast quantities of data, enabling police to trace individuals associated with crime or terrorism. Such systems infringe privacy since individuals may be monitored and judged without the court's permission.[21] The imprecise manner in which AI determines what a security threat is gives individuals concern about wrongful surveillance, data privacy, and human rights abuses.[22]

Crime and Criminal Tracking Network & System, or CCTNS, is a unified database that is based on AI to store and track criminal records, automate police reports, and facilitate inter-agency coordination.[23] Since CCTNS is based on old crime data to assess risk, it may perpetuate existing biases in the justice system.[24] If AI algorithms unjustly tag certain communities as high-risk based on flawed data, it may result in wrongful arrests, discriminatory policing, and the loss of public trust in the police.[25]

The Automatic Facial Recognition System is used by the National Crime Records Bureau to assist the police in tracking suspects via CCTV cameras and image databases.[26] The system tends to make mistakes, especially in identifying marginalized community individuals.[27] AI face recognition mistakes have resulted in false arrests and violations of privacy.[28] Moreover, no policy on how the data is to be stored, accessed, or utilized is set.[29]

---

[19] Supreme Court E-Committee, 'Report on CMS Deployment' (2021) https://ecourts.gov.in/ accessed 18 February 2025.
[20] NATGRID, 'AI and Surveillance: Enhancing National Security' (2023) https://www.natgrid.gov.in/ accessed 18 February 2025
[21] Gautam Bhatia, Privacy and AI Surveillance in India (OUP 2022) 89.
[22] Supreme Court of India, Justice KS Puttaswamy v Union of India (2017) 10 SCC 1
[23] NCRB, 'Crime and Criminal Tracking Network & System (CCTNS): AI in Policing' (2022) https://ncrb.gov.in/ accessed 18 February 2025
[24] Rashmi Raman, 'AI in Policing: Bias and Discrimination in CCTNS' (2023) 14(1) Indian Journal of Criminal Law 55
[25] Pratiksha Baxi, 'Criminal Justice and Algorithmic Decision-Making in India' (2023) 10(2) Indian Law Review 72.
[26] NCRB, 'Automatic Facial Recognition System: A Critical Assessment' (2023) https://ncrb.gov.in/ accessed 18 February 2025
[27] Smriti Parsheera, 'Facial Recognition and Privacy: Legal Concerns in India' (2023) 9(3) NLU Delhi Journal of Technology and Law 148
[28] Anirudh Burman, 'AI in Policing: The Ethical and Legal Challenges of AFRS' (2023) 11(1) Journal of Cyber Law 87
[29] Vidhi Centre for Legal Policy, 'AI and the Law: A Regulatory Roadmap for India' (2023) https://vidhilegalpolicy.in/ accessed 18 February 2025

There has to be trust, but there is accountability too. If an AI system decides something in a biased or unfair manner, it will be difficult to locate the bias and whom to hold accountable. In the absence of transparency, it is difficult to ensure that things are level and discrimination is kept at bay. This absence of accountability and laws to govern AI will have far-reaching consequences on the Indian legal system.

REVIEW OF LITERATURE

Yavar Bathaee (2018) – "The Artificial Intelligence Black Box and the Failure of Intent and Causation:" The author of this paper explores how deep learning models operate as Black Boxes that make it difficult to determine the intent and causation in their outputs. Traditional legal doctrines of intent (mens rea) and causation (liability principles) are ineffective for AI or deep learning systems. The author argues that the legal system is ill-equipped to regulate AI-based decision-making, as the law has historically been structured to make humans accountable rather than autonomous systems. The paper suggests that new legal frameworks must be developed to ensure accountability and explainability for AI systems.

Sandra Wachter et al. (2018) – "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR:" In this paper, the author proposes counterfactual explanations against the lack of transparency in AI decision-making. Counterfactual explanations make humans understand how different inputs can lead to a different decision, without having access to the algorithm or internal logic of that AI system. The author argued that this method is useful under the General Data Protection Regulation (GDPR), which grants humans the right to explanation when subjected to AI-based decision-making. The paper further suggests the method of counterfactual explanations as an alternative to enhance accountability without compromising the intellectual property rights of the AI developers.

Frank Pasquale (2015) – "The Black Box Society: The Secret Algorithms That Control Money and Information:" The author in this book highlights the secrecy of AI decision-making systems that limit transparency in fields like finance, healthcare, and governance. The author stated that the Black Box of AI enables corporations and government institutions to make impactful decisions without any accountability. The study further emphasizes the dangers of opaque algorithms in shaping economic and social outcomes with a very little overview of how these systems arrive at their decisions. The author warns that if these AI systems are

allowed to operate without any supervision in legal systems, it could lead to unfair and unchallengeable decisions that violate the principle of the rule of law.

Janees Rafiq (2024) – "Harnessing the Power of Artificial Intelligence in Indian Justice System: An Empirical Study:" The author of this paper explores the potential of AI in the context of India's legal framework. The paper evaluates how AI can address certain challenges like case backlogs and procedural delays, streamline the tasks performed by lawyers, and aid judges in the decision-making process to reduce the overall duration of trials. The author further assessed existing AI applications in the Indian justice system by examining their effectiveness in enhancing judicial efficiency and transparency. The study also highlights the ethical considerations and the need for proper legal frameworks to govern the use of AI in judicial processes.

Sahil Girhepuje (2023) – "Are Models Trained on Indian Legal Data Fair?:" The authors in this paper investigate bias and fairness in AI models trained on Indian legal data. The authors find that AI models trained on Indian court decisions carry certain stereotypes like socio-economic and caste-based disparity, which lead to potentially unfair bail recommendations. The authors in this paper highlight the need for ethical consideration and research requirements to understand the fairness and bias issues in the AI models in the legal system. The paper further stresses mitigating the bias in these AI models through an effective algorithmic approach to ensure that these models are supported effectively and fairly.

EMERGENCE OF LEGISLATIONS GOVERNING AI: AN INTERNATIONAL PERSPECTIVE

"*The development of full artificial intelligence could spell the end of the human race.*"

– Stephen Hawking.

The mass adoption of AI tools by various sectors, including legal systems around the world, for efficiency and effective decision-making, poses significant risks if left unregulated. The lack of transparency in AI systems or the Black Box problem raises various concerns, including bias, accountability, and ethical compliance. It has encouraged governments and international organisations to develop legislative frameworks to address the legal, ethical, and societal

challenges posed by these AI systems by regulating the design, deployment, and use of these systems.[30]

European Union Artificial Intelligence Act 2024 (EU AI ACT 2024): This act is the first-ever legal framework on AI drafted by the European Union, a comprehensive framework that deals with the risks of Artificial Intelligence. The Act aims to ensure the human-centric and ethical development of artificial intelligence. The Act defines 4 levels of risk for AI systems, with each level requiring a different degree of regulation.

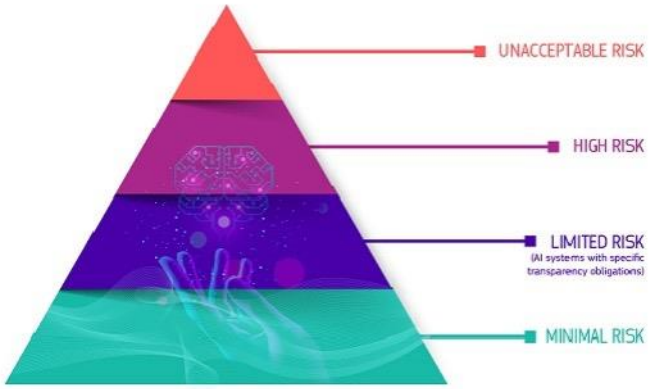

Figure 2: Levels of Risk

Minimal Risk: This level includes the majority of AI programs, such as AI-enabled video games and spam filters. These programs are safe; hence, they don't need any regulation.[31]

Limited Risk: AI systems in this category include deepfakes and chatbots like Chat GPT, Gemini, etc. The main rule for AI systems is to be transparent. The Act introduces obligations for AI systems to ensure that humans are informed when necessary to preserve trust. For instance, when using AI systems such as chatbots, humans should be made aware that they are interacting with a machine so they can make an informed decision.[32]

High Risk: This level covers AI in critical areas where decisions and actions can have a profound impact on people's health, safety, or fundamental rights. These include:

[30] Savio Jacob, "AI Regulations Around the World: A Comprehensive Guide To Governing Artificial Intelligence" (2024). Retrieved from https://www.spiceworks.com/tech/artificial-intelligence/articles/ai-regulations-around-the-world
[31] European Union, "European approach to artificial intelligence", (2024) Retrieved from https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence accessed 10 February 2025.
[32] Ibid

- AI safety components in critical infrastructure.
- AI systems used for remote biometric identification;
- AI solutions are used in the administration of justice and democratic processes.[33]

For instance, in transportation, AI-powered self-driving cars make decisions in a fraction of a second to keep passengers safe on the road. In healthcare, where AI can assist in surgeries and diagnosis, where a small mistake could have major consequences. In the education system, AI can be used to grade exams and assess student performance. The AI system should be fair and must avoid bias, with human oversight ensuring just outcomes. [34] AI systems at this level are subject to strict obligations and must undergo thorough risk assessment, use top-quality data, maintain detailed logs, and always have human oversight.

Unacceptable Risk: Those AI systems that are a threat to the safety, livelihood, and rights of the people are banned in the European Union. The AI Act prohibits AI systems that engaging in the practice:

- Harmful manipulation and deception;
- Harmful exploitation of vulnerabilities;
- Social Scoring;
- Individual crime offense risk assessment or prediction;
- Emotion recognition in workplaces and educational institutions.

The EU AI Act 2024 implemented a structured legal framework to regulate AI systems by categorising them according to risk level to ensure transparency, accountability, and human oversight in these systems. However, the success of this act will depend on the effective enforcement and adaptation to evolving AI challenges.

UNITED STATES: ALGORITHMIC ACCOUNTABILITY ACT 2023 & AI BILL OF RIGHTS 2022

The Algorithmic Accountability Act and AI Bill of Rights represent the United States' efforts to promote accountability, transparency, and fairness in AI systems. Algorithmic Accountability is a proposed law that requires companies deploying AI in various sectors, including legal systems, healthcare, and finance, to be transparent about the algorithms they

[33] ibid
[34] *Ibid*

use and to ensure that their AI systems are fair and unbiased. However, the compliance with this Act is largely voluntary, as it lacks strict enforcement mechanisms and does not impose mandatory transparency standards. [35]

The AI Bill of Rights is a policy framework rather than a law. It is intended to support the development of policies and practices that protect civil rights and promote democratic values in the deployment and governance of automated systems.[36] It consists of five core principles to help guide the design, use, and deployment of AI systems, which include:

- Safe and effective AI systems;
- Algorithmic discrimination protections;
- Data Privacy;
- Explainability in AI systems.
- Human oversight.[37]

Recognizing the growing concerns and lack of strict enforcement around AI governance, the U.S. government established the US AI Safety Institute to understand, identify, and mitigate the risk of advanced AI systems to harness the enormous potential of this technology and the risk they pose for public safety and national security.

CHINA: AI GOVERNANCE THROUGH STATE-CONTROLLED REGULATION

China has been a particularly proactive actor in designing and creating some of the earliest laws on AI. In 2021, China promulgated a comprehensive set of rules concerning algorithms, deepfakes, and generative intelligence services. It also imposes a licensing regime for introducing generative AI services. AI systems in China must align with national policies and go through several inspections and approvals before they can be deployed. This centralised regulation ensures that AI systems must follow the state's ideological and political values, which can limit transparency and innovation in various sectors, particularly in the legal system.[38]

[35] National Institute of Standards and Technology (NIST), 'AI Risk Management Framework' (2023) https://www.nist.gov/ accessed 11 February 2025.
[36] Tom Krantz, "What is the AI Bill of Rights?" (2024). Retrieved from https://www.ibm.com/think/topics/ai-bill-of-rights
[37] Ibid
[38] James Gong, "Ai Governance in China: Strategies, Initiatives, and Key Considerations", (Bird& Bird, 2024) Retrieved from https://www.twobirds.com/en/insights/2024/china/ai-governance-in-china-strategies-initiatives-and-key-considerations

The compliance with AI and digital content in China is structured around various key pillars such as:

- Content Moderation;
- Data Protection;
- Algorithmic Governance.

Pillar	**Legal Instrument**	**Key consideration**
Content Moderation	• 2017 CSL • Administrative Measures for Internet-based Information Services 2011 • Provisions on the Governance of Network Information Content Ecolog 2019 • 2022 Deep Synthesis Provisions • 2023 Generative AI Measures	The content generated through the AI systems must align with the state-approved narratives. It requires labels to identify the AI-generated content, security assessment, and transparency. It also establishes internal rules for identifying and taking down illegal content.
Data Protection	• 2021 PIPL • 2021 DSL • 2021 Recommendation Algorithm Provisions • 2022 Deep Synthesis Provisions • 2023 Generative AI Measures	The personal information of the user must be obtained with the user's consent and organizations must have a lawful basis for such use of information. The companies must adhere to the principles of legality, sincerity, transparency, and accountability.
Algorithmic Governance	• 2021 Recommendation Algorithm Provision	It is an emerging pillar of Chinese regulation of AI

	• 2022 Deep Synthesis Provisions • 2023 Generative AI Measures • Measures for Ethical Review of Science and Technology (Trial)	systems. The regulation will vary for different algorithms but the companies must ensure that these systems do not discriminate, go through security assessments, and share algorithmic information with the authority.

OECD (ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT) AI PRINCIPLES:

OECD AI principles are the first intergovernmental standards on AI that establish various guidelines for the development and use of AI technologies responsibly.[39] The framework for governance is primarily focused on the ethical use of these systems along with accountability and transparency of AI systems. [40]

These principles advocate the use of AI systems that are innovative and trustworthy and adhere to human rights and democratic values. Some of these value-based principles are:

- Inclusivity and Human-centric values;
- Fairness, Transparency, and Explainability;
- Robustness and safety;
- Accountability.[41]

It further provides recommendations for policymakers to invest in AI research and development, international cooperation for trustworthy AI, and regulation of AI systems with

[39] OECD, 'OECD Principles on Artificial Intelligence' (OECD, 2019) https://www.oecd.org/going-digital/ai/principles/ accessed 12 February 2025.
[40] OECD, 'Recommendation of the Council on Artificial Intelligence' (2019) https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 accessed 12 February 2025.
[41] Ibid

a risk-based approach. While these principles are not binding upon nations, they provide valuable frameworks that can shape policy development worldwide.[42]

UNESCO'S RECOMMENDATION ON THE ETHICS OF ARTIFICIAL INTELLIGENCE (2021)

UNESCO (United Nations Educational, Scientific and Cultural Organisation) has produced the global standard on AI ethics, which is applicable to all 194 member countries of UNESCO.[43] It sets ethical guidelines to ensure that AI systems align with human rights, transparency, fairness, and accountability. The recommendations outlined the key ethical principles for AI systems, which include:

- Human Rights and Dignity;
- Fairness and Non-Discrimination;
- Transparency and Explainability;
- Sustainability and Environmental Responsibility;
- Accountability and Governance.[44]

To ensure that AI systems are developed and used ethically, these recommendations introduce several compliance measures to ensure that AI aligns with human rights, fairness, and social inclusion. AI systems must protect privacy and ensure data security to prevent misuse. Countries should create laws to promote transparency and accountability in AI systems.[45]

UNESCO has also provided various tools such as the readiness assessment methodology, ethical impact assessment tool, and AI ethics observatory to help countries implement these policies. However, the enforcement relies on the national government, as this framework is not legally binding. Strong enforcement and international collaboration are required to transform these principles into binding regulations to achieve the objective of these legislations.

[42] OECD, 'Governance of Artificial Intelligence: Who is Responsible?' (2021) https://www.oecd.ai/en/governance accessed 12 February 2025.
[43] UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (2021) https://unesdoc.unesco.org/ark:/48223/pf0000381137 accessed 20 February 2025.
[44] UNESCO, 'AI Ethics: Global Standards for Responsible AI' (2022) https://www.unesco.org/en/artificial-intelligence/recommendation-ethics accessed 20 February 2025.
[45] UNESCO, 'Readiness Assessment Methodology for Ethical AI' (2022) https://www.unesco.org/en/artificial-intelligence/assessment-tools accessed 20 February 2025.

CONVENTION 108+ AND THE GENERAL DATA PROTECTION REGULATION (GDPR)

Convention 108+ and the GDPR are significant international legal frameworks that are shaping AI governance and data protection by safeguarding the personal data of individuals. However, they vary in application, scope, and legal effect. Convention 108+, introduced by the Council of Europe, provides a broad framework for the member states to develop national legislation based on data protection and to promote transparency, accountability, and fundamental rights of individuals in AI-driven decision-making. In contrast, the GDPR, a directly applicable regulation in the European Union and the European Economic Area, provides extensive rights to individuals and imposes strict obligations on organisations handling personal data.

Article 22(1) of the GDPR protects individuals subjected to automated decisions by AI systems if these decisions have legal or similarly significant effects. It ensures individuals have the right to human intervention, to contest decisions, and to receive an explanation.[46] GDPR imposes heavy penalties for non-compliance that make it the most robust AI governance framework in the world. While Convention 108+ outlines general principles, the GDPR outlines strict accountability and transparency requirements for the effective enforceability of these laws.[47]

EMERGENCE OF LEGISLATION GOVERNING AI IN INDIA

Presently, India does not have a dedicated AI regulation, and the existing legislative framework primarily governs cybersecurity, digital data protection, criminal law, and electronic evidence. Various statutes indirectly regulate AI-related aspects, particularly in data protection, digital forensics, surveillance, and electronic records. However, these laws lack explicit provisions addressing AI decision-making, transparency, and accountability, creating regulatory gaps in AI-assisted judicial and law enforcement applications. India has a complex legal system of various laws, such as the Constitution, statutory laws, rules, regulations, and guidelines. The common view across the state, organisations, and civil society groups is that AI risks can be addressed through existing legal frameworks.

The table below illustrates how existing statutory laws can deal with the risks of AI systems.

[46] GDPR 2024, Article 22(1)
[47] Janees Rafiq, "An Examination of Right to Privacy as a Fundamental Human Right in India & UK: A Comparative Analysis" (Unpublished).

Nature of Harm	**Applicable Statutory Laws**
Depiction of a child in a sexually explicit video that is AI-generated	• Information Technology Act, 2000 • Prevention of Children from Sexual Offences Act, 2012 • Bharatiya Nyaya Sanhita, 2023
Unauthorized impersonation using AI-generated deepfakes	• Bharatiya Nyaya Sanhita, 2023 • Information Technology Act, 2000
Discrimination in hiring decisions using AI recruitment tools	• Rights of Persons with Disabilities Act, 2016 • Transgender Persons (Protection of Rights) Act, 2019 • Code on Wages, 2019 • Scheduled Castes and the Scheduled Tribes (Prevention of Atrocities) Act, 1989
Use of an individual's personal data without consent to train AI models	• Digital Personal Data Protection Act, 2023 • Information Technology Act, 2000
Misleading advertisements about the reliability or performance of an AI service	• Consumer Protection Act, 2019
Use of copyright-protected material in AI-generated content without permission	• Copyright Act, 1957
AI-driven surveillance and data collection for national security or law enforcement	• Indian Telegraph Act, 1885 • Information Technology Act, 2000 • Bharatiya Nyaya Sanhita, 2023

Automated decision-making leads to harm in financial or healthcare sectors.	• Consumer Protection Act, 2019 • Digital Personal Data Protection Act, 2023 • The Contract Act, 1872
AI-powered facial recognition used without consent for surveillance	• Indian Telegraph Act, 1885 • Information Technology Act, 2000

UNPRECEDENTED ISSUES AND CHALLENGES POSED BY THE GROWTH OF AI AND BLACK BOX

The integration of Artificial Intelligence into the legal system raises concerns like transparency and due process, as well as accountability and fairness. The AI systems provide efficiency and data-driven insights, but the Black Box nature of these systems presents serious issues when deployed in the legal system. Many AI applications, including predictive policing and risk assessment tools, together with automated sentencing models, function without revealing their logical processes, making them challenging for regulators.

The traditional legal principles struggle to adapt to AI decision-making systems because deep learning systems produce untraceable links between data inputs and their resulting outputs, which conflict with current legal standards on intent and causation.[48] AI models both focus power on those who write the algorithms while cutting out public oversight.[49] Without standardised legal rules guiding AI-assisted legal decisions, these risks become more severe. Although courts have started to consider AI-related issues like facial recognition system biases, along with unexplained algorithmic decisions, they lack specific laws to guarantee AI tool transparency and equity in legal applications. In judicial systems, the Black Box represents a major problem because AI systems cannot explain their actions, which threatens fundamental rights such as due process and the right to a fair trial. This section reviews the problems created

[48] Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation" (2018) 31 Harvard Journal of Law & Technology 889
[49] Frank Pasquale, "The Black Box Society: The Secret Algorithms That Control Money and Information" (Harvard University Press 2015)

by AI within legal systems through issues like opacity, along with algorithmic bias and gaps in accountability, all while potentially breaking constitutional guarantees.

Former Chief Justice of India, D.Y. Chandrachud, articulated that AI confronts manifold risks even in judicial decision-making: "We are conscious of the fact that artificial intelligence has a flip side. For instance, it would be very difficult and difficult for us to allow artificial intelligence to determine conviction in criminal cases. I don't think that we, as judges, would ever want to see the discretion we exercise, based on sound judicial principles, being overtaken by AI in sentencing policy."

This statement highlights the judiciary's reluctance to rely entirely on AI, as it lacks accountability, transparency, and human reasoning. This section will analyse the critical issues posed by AI in the legal system, particularly in the Indian context.

Lack of Transparency and Accountability: An underlying problem with AI systems is that they are not transparent about their algorithms. These are invariably complex machine learning models trained on vast data sets that are difficult to interpret. The absence of transparency regarding how these systems arrive at their conclusions contributes to the Black Box nature of many AI-driven decisions in criminal justice. The opacity built into these artificial intelligence systems blocks defendants, judges, and legal experts from seeing or challenging the reasoning behind pivotal choices, such as bail denials, sentencing suggestions, and parole decisions.

The most unnerving feature of these algorithms is that they cannot provide human-readable explanations of their predictive results. Judges in traditional judicial decision-making typically balance a combination of various factors, such as individual facts, evidence, and legal rules. AI systems, on the other hand, employ computational algorithms operating on large databases without a general understanding of the rationale behind specific outputs. For example, when a predictive model is applied to determine whether an individual will re-offend, the system will return a risk score for that. Yet, the basis of that score can be on a set of variables, and some of those could be challenging for even the person who designed the system to entirely explain. This unexplainability can result in an unreliable judicial environment where such important decisions are taken without sufficient scrutiny.

Impact on Judicial Decision-Making and Public Trust: The 2016 ProPublica investigation highlights the problematic nature of such predictive algorithms, which can yield radically flawed results that disproportionately affect marginalised populations, most significantly racial

minorities.[50] The results depicted in ProPublica's report identified one particular algorithm widely used in U.S. courts to predict defendants' recidivism and found it to be biased against Black defendants. Specifically, the report illustrated that Black defendants were almost twice as likely to be inappropriately predicted to re-offend when they did not, compared to White defendants, who were disproportionately misclassified as low-risk. This sharp racial disparity highlights the potential of embedded historical biases in training data to lead to discriminatory conclusions, which are hard to challenge or even locate due to the lack of transparency in the operational mechanisms of these algorithms. In a judicial system that is meant to provide justice and equality to everyone, a lack of transparency in decision-making is a negative factor in public trust in the system. When people cannot understand the rationale behind some decisions or cannot challenge them sufficiently, they begin questioning the validity of the judicial process.

ALGORITHMIC BIAS AND DISCRIMINATION

AI systems are programmed with a series of algorithms and learn from data to discern patterns. They are therefore susceptible to both biases in the algorithms used, as various groups of engineers bring extremely different biases and assumptions to the design of algorithms and the datasets.[51] Various legal AI systems function with various algorithms and, in the majority of cases, on different datasets. It is therefore expected that different legal AI systems can have different outcomes. Systematic and unfair discrimination by the algorithms can perpetuate or provoke existing social inequalities. These algorithms are used to predict outcomes with historical data that can inherit racial, gender, or socioeconomic disparities that can lead to unintentional bias and can further impact the fairness of the legal system.

The root cause of Algorithmic Bias:

Data Bias: Prejudices are reflected in historical datasets that are frequently used to train algorithms. An AI system used for risk assessment in criminal justice is likely to reproduce and even magnify these biases if it is trained on arrest records that disproportionately target

[50] Julia Angwin and others, "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." (2016), Pro Publica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[51] Ronald Yu, "What's Inside the Black Box? AI Challenges for Lawyers and Researchers.", (2019), Article in Legal Information Management.

minority communities. The issue with using historical data is that it naturally contains systemic inequality patterns, which the machine then learns.

Bias in Algorithm Design: Engineers' design decisions can cause algorithms to produce biased predictions even when they don't specifically use demographic variables like gender or race. These decisions may involve the selection of features, the analysis of data, and the predictions that the algorithm has the capacity to make.

Overfitting and Underfitting: Overfitting is a phenomenon in machine learning where a model becomes too specific to a particular dataset and is unable to generalise. If an overfitted algorithm is applied to a biased dataset, it will amplify those biases. Underfitting happens when an algorithm is too basic and is unable to pick up on complex relationships within the data, and may even miss key variables that would result in a more accurate and equitable prediction.

Algorithmic Bias in the Legal System: Case Studies and Analysis:

COMPAS: The Correctional Offender Management Profiling for Alternative Sanctions is perhaps the most applied risk assessment tool in the US. It helps judges forecast a defendant's likelihood of recidivism and guide bail, parole suitability, and sentencing recommendations. In 2016, investigative reporters at ProPublica conducted an in-depth examination of COMPAS by studying more than 7,000 criminal cases in Broward County in Florida. Their results showed that the algorithm inappropriately gave higher risk scores to Black defendants than to white defendants, even after adjusting for their criminal histories and backgrounds. The research discovered that Black defendants were twice as likely to be misclassified as high risk, while white defendants were more likely to be misclassified as low risk.[52]

HARM: The Harm Assessment Risk Tool is an AI risk assessment system that has been created by the Durham Constabulary in the UK. It classifies persons as low, moderate, or high risk to re-offend according to machine learning algorithms based on historical police data. A Big Brother Watch UK study discovered that HART was racially and socioeconomically biased, overclassifying people from low-income communities as high-risk. The tool was based on police records that captured past over-policing of some communities, creating self-perpetuating biases. For instance, communities with traditionally higher arrest rates—frequently the result

[52] Julia Angwin, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks" (2016) Pro Publica. Accessed https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing accessed 15 February 2025

of proactive policing rather than higher crime rates- were disproportionately represented in HART's forecasts.[53] In addition, the transparency of the algorithm was an issue since defendants did not have access to the entirety of how the risk scores were calculated. Such a lack of transparency created potential breaches of due process under the European Convention on Human Rights, most notably Article 6 (fair trial). [54]

PredPol (Predictive Policing), a system employed by police agencies throughout the United States, examines crime statistics to forecast areas most prone to crime, which police departments then allocate patrols based on. Yet various studies have shown pronounced racial and geographic biases within PredPol's algorithm. A 2019 analysis by the Human Rights Data Analysis Group (HRDAG) discovered that PredPol disproportionately sent police to areas with more Black and Hispanic residents, even when crime was lower than in predominantly white areas.[55] This was because crime data that was historically used to train PredPol's model already contained biased policing patterns, and thus, a feedback loop was created in which already over-policed communities were further over-policed.[56]

In addition, a 2021 MIT Technology Review report discovered in an investigation that PredPol's crime predictions disproportionately targeted low-level offenses in disadvantaged neighbourhoods and overlooked white-collar offenses. This absence of data created fears that, far from responding to real patterns of crime, predictive policing was perpetuating structural injustice within the legal system.[57]

These examples demonstrate the risks of algorithmic bias within the justice system, especially where AI technology is used in sentencing, policing, and law enforcement surveillance. Many AI tools function as black boxes, such that defendants have no way to contest or grasp algorithmic outcomes. AI systems that are trained on biased historical data tend to reproduce and reinforce the existing discrimination. Biased AI systems are a concern for constitutional and human rights issues, especially for due process and equal protection under the law.

---

[53] Big Brother Watch, 'Predictive Policing and the HART Algorithm' (Big Brother Watch, 2021) https://bigbrotherwatch.org.uk/ accessed 15 February 2025
[54] European Convention on Human Rights 1950 -Article 6.
[55] Kristian Lum And Tarak Shah, "Measures Of Fairness For New York City's Supervised Release Risk Assessment Tool" (2019), Human Rights Data Analysis Group
[56] Sarah Brayne, "Predict and Surveil: Data, Discretion, and the Future of Policing" (Oxford University Press 2021)
[57] Karen Hao, 'What's Wrong with AI Crime Prediction? Everything' (MIT Technology Review, 2021) https://www.technologyreview.com/2021/01/07/1016221/whats-wrong-with-ai-crime-prediction-everything/ accessed 15 February 2025

In response, several regulatory and legal measures have been put forward, including the EU Artificial Intelligence Act and more robust data protection rules under GDPR, to increase transparency and accountability in AI-driven decisions. Judicial checks and policy changes are needed on an ongoing basis to ensure AI does not aggravate systemic inequalities instead of reducing them.

Privacy and Data Breach: The use of Artificial Intelligence in legal systems has raised issues regarding privacy and data protection. AI technologies operate based on huge databases, comprising personal data like biometric data, criminal records, and case files. These technologies introduce profound legal and ethical issues around the data gathering, storage, and utilisation procedures.

One key concern is the fact that AI systems process personal data without explicit user consent or the provision of effective protective means, thereby inviting risks of unsanctioned surveillance and information leaks. The majority of the AI tools are black boxes, which are not transparent, hence making it difficult for people to comprehend how their information is processed.

This risk becomes extremely perilous in predictive policing systems, face recognition systems, and digital court case management platforms in judicial systems. The potential of AI in collecting and processing large volumes of judicial data makes it an inevitable tool for police forces and court institutions. Nevertheless, the same capability introduces immense privacy concerns owing to the power of mass surveillance and possible data abuse. The lack of effective supervisory regulation further worsens the dilemma, as many jurisdictions are yet to establish well-defined legal frameworks to address AI-driven data processing in the justice system.

Privacy Issues in AI-Driven Legal Systems: The application of AI in legal systems entails the processing and collection of vast amounts of personal data, leading to significant privacy concerns. One of the primary issues is the application of mass data surveillance by AI systems, monitoring individuals without explicit consent. The Metropolitan Police in the United Kingdom has rolled out Live Facial Recognition technology in public areas, scanning hundreds of faces in real time to detect suspects. The practice has been criticised across the board for infringing on people's rights, as enshrined in the European Convention on Human Rights, Article 8.[58] Another important privacy issue is the non-transparency of AI algorithms, or the

[58] R (Bridges) v South Wales Police [2020] EWCA Civ 1058.)

black box problem. AI systems base their decisions on patterns or trends in data, but users like legal professionals or concerned individuals are not able to understand the rationale of AI systems. This transparency problem makes it difficult for individuals to challenge the legality of AI-based decisions because they are not completely aware of the reasons behind them.

In addition, AI systems can process data without the explicit consent of the individuals. The majority of AI systems run through arrangements among the government, the private sector, and the tech industry, which poses the question of whether personal data is ethically harvested. The controversial case of Clearview AI scraped billions of social media photos without authorisation from their owners. Court cases filed against Clearview AI allege that its operations are in violation of data protection acts and the right to privacy.[59]

In India, Automated Facial Recognition Systems have sparked grave privacy issues. The National Crime Records Bureau had suggested AFRS as an instrument for real-time suspect identification through CCTV surveillance. The system is aimed at boosting national security, but is criticised for having the potential to infringe on privacy rights based on its mass surveillance capacity.

Legal scholars contend that AFRS does not have a strong legal framework since there is no data protection law, let alone all-encompassing in nature, that specifically governs its use. There have also been worries about racial and religious profiling since research indicates that face recognition technology based on AI racially misidentifies individuals from ethnic minority groups. Also, the application of AFRS by omission of judicial control and express parliamentary sanction makes it prone to abuse.[60]

A report by the Internet Freedom Foundation (IFF) described how AFRS can potentially be against the Supreme Court judgment in the K.S. Puttaswamy case that recognized privacy as a fundamental right.[61] In the absence of guidance on data retention, access control, and responsibility mechanisms, AFRS seriously threatens civil liberties in India.[62]

[59] Kashmir Hill, 'The Secretive Company That Might End Privacy as We Know It' (The New York Times,2020) https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html/ accessed 15 February 2025
[60] Sameer Yasir, 'India's Surveillance State', (2021) The New York Times.
[61] Justice K.S. Puttaswamy v Union of India (2018) 10 SCC 1
[62] Kashmir Hill, 'The Secretive Company That Might End Privacy as We Know It' (The New York Times,2020) https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html accessed 15 February 2025

THE WAY FORWARD: SOLUTIONS & POLICY RECOMMENDATIONS

We have regulations in healthcare education and financial services, but next to none in AI, even though it's such a large and growing aspect of human life, particularly in the legal system. We are aware of the digital utopia that AI can provide us with. So, we should introduce regulations to ensure we reach this utopian situation and avoid a dystopian one.

Explainable AI (XAI): As described earlier, since AI learns from the surrounding background and previous errors, even programmers struggle to understand the machines' internal logic and decision-making. In this context, calls for an enhanced understanding of AI's inner processes are ubiquitous. The growing application of AI in the legal system has raised profound concerns about transparency, accountability, and trust. AI-powered tools are increasingly being utilized in case law analysis, predictive sentencing, legal research, and risk assessment, but their non-transparent decision-making processes, usually termed the Black Box, are raising challenges to the rule of law. Explainable AI (XAI) solves this problem by understanding the basis of the decision and the opportunity for individuals to challenge its decisions. XAI allows judges, attorneys, and parties to the action to understand and contest AI-made suggestions, bringing the applications of AI in line with legal requirements of due process, fairness, and responsibility.

Researchers over the years have developed various XAI frameworks that can improve transparency in AI systems. In pre-model training of AI systems, use an interpretable or simple model design rather than a complex or deep learning black box. The logic of these systems should be designed as per the legal reasoning principles and ensure that AI systems do not impact any group of society disproportionally. For complex AI systems, post-hoc explanations can help provide interpretable explanations. Through feature importance analysis, key variables can be identified that influence the system's decision. By using Local Interpretable Model -Agnostic Explanations or LIME, which creates simple human human-interpretable models of AI systems.[63] Shapley Additive explanations, or SHAP, can produce consistent explanations by calculating the contribution of each feature of the algorithm that is used to predict the outcome. [64] Explainable AI is important and can enhance the role of AI in the legal

[63] Ribeiro, Singh, and Guestrin, 'Why Should I Trust You? Explaining the Predictions of Any Classifier' (2016) Proceedings of the 22nd ACM SIGKDD 1135.
[64] Scott M Lundberg and Su-In Lee, 'A Unified Approach to Interpretable Machine Learning with SHAP' (2017) Advances in Neural Information Processing Systems (NeurIPS) 4765

system by making sure that these systems are interpretable, transparent and fair while making decisions on their own and do not work as black boxes.

Algorithm Audits: Algorithm Auditing is an important mechanism that ensures that AI systems remain transparent, fair, and free from all biases while making decisions on their own. It involves systematic reviews to detect and prevent biases before deploying AI systems and their use.[65] Auditing includes the testing of AI models for biases that are found on a variety of factors such as race, gender, caste, or socioeconomic factors. It is required to test the models for accuracy and reliability in the legal system, to ensure the rights of the individuals and the rule of law.[66]

Algorithm Auditing can be implemented by pre-deployment audits that can be conducted before the AI models are deployed or through regular assessment after deployment to detect bias that may emerge over time due to the evolving trends in data.[67] Algorithmic Impact Assessments should be conducted to ensure that AI models do not end up causing disproportionate harm to any section of society.[68] Just like Environmental Impact Assessments (EIA) work in environmental law, AIAs would be for measuring the impact of AI systems on society prior to deployment. The corporations should be required to publish evaluation reports for bias, as it is mandated under the GDPR. These reports should cover all the methodologies that are used to design the logic of AI systems, test for bias and the actions taken to mitigate the bias in AI systems.[69] The combination of these approaches would ensure public confidence in AI tools through ensuring observance of constitutional and human rights requirements, thus upholding the rule of law in the legal system.

Human in the Loop System: The concept of the HITL AI system introduces a hybrid model where the decisions made by AI remain subjective to human oversight and intervention.[70] It ensures that humans retain ultimate control over all the decisions made by AI systems to reduce the risk posed by these opaque and potentially biased algorithms.[71] Fully automated AI systems

[65] Joshua A Kroll et al, 'Accountable Algorithms' (2017) 165 University of Pennsylvania Law Review 633.
[66] Frank Pasquale, "The Black Box Society: The Secret Algorithms That Control Money and Information" (Harvard University Press 2015)
[67] Cary Coglianese and David Lehr, 'Algorithmic Regulation and Transparency' (2020) Harvard Journal of Law & Technology 23
[68] AI Now Institute, 'Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability' (2018) https://ainowinstitute.org/aiareport2018.pdf
[69] European Commission, 'The AI Act: A Regulatory Framework for Trustworthy AI' (2024)
[70] Cary Coglianese and David Lehr, 'Transparency and Algorithmic Governance' (2019) 71 Admin L Rev 1.
[71] Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (Harvard University Press 2015).

inherit risks that are potentially for errors, biases and a lack of contextual understanding, which can lead to unfair legal rulings.[72] In cases like State v Loomis, where AI-driven risk assessment tools were employed at sentencing, the lack of transparency and human intervention raised constitutional concerns.[73] Human-in-the-loop (HITL) systems are meant to bridge the gap between the efficiency of AI systems and the legal safeguards required by ensuring that AI-suggested recommendations always go through human review before they can affect judicial rulings.[74]

Continuous Human Review and Intervention is one of the components of an effective HITL system, where the AI decisions must be reviewed and validated by humans before they influence the final decisions.[75] For instance, in predictive justice, where AI may suggest the decision for bail or parole but the final ruling should be made only by a judge. AI tools that are used for contract analysis, legal research, and case prediction must serve as assistive tools rather than decision-makers.[76]

Human feedback should be incorporated into AI models to correct discriminatory patterns and AI biases, as AI might overlook the case-specific details and cannot interpret the broader context of legal cases, unlike humans. AI models should be designed to learn from decisions made by humans and hence refine future decisions based on legal expert feedback.[77]

LEGAL FRAMEWORK FOR AI GOVERNANCE IN INDIA

Legal frameworks are an essential and effective solution to the black box problem in the legal system. Regulatory mechanism safeguards constitutional principles, procedural fairness, and public trust in an AI-assisted legal system by ensuring transparency, accountability, and fairness in its decision-making.[78]

Indian legal system currently does not have any dedicated laws for AI and the black box problem, instead relying on the general laws such as the IT ACT 2000, DPDP Act 2023 and BNSS 2023. These laws are not sufficient for governing AI in the Indian legal system. Without

[72] Solon Barocas and Andrew Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671.
[73] State v Loomis 881 N.W.2d 749 (Wis. 2016).
[74] Ibid
[75] AI Now Institute, 'Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability' (2018) https://ainowinstitute.org/aiareport2018.pdf
[76] European Commission, 'The AI Act: A Regulatory Framework for Trustworthy AI' (2024).
[77] UK Ministry of Justice, 'AI in Judicial Decision-Making' (2024)
[78] Frank Pasquale, "The Black Box Society: The Secret Algorithms That Control Money and Information" (Harvard University Press 2015).

proper AI laws, the system is being operated in a grey area, which increases the risk of violating an individual's rights.[79]

The judiciary should come up with rules and procedural safeguards to regulate the use of AI in the legal system. [80]AI tools must not replace judicial discretion but should function as an assistive tool. There should be mandatory explainable measures where AI models must provide clear and interpretable justification for their decisions. An AI system should not be allowed to make the final decision without human validation.

A regulatory body should be established to evaluate and monitor the compliance of AI in the legal system by regularly conducting Audits to detect and mitigate bias or errors in the AI models. The body must ensure that only validated and compliant AI systems are being deployed in the legal system. For non-compliance or unethical deployment of AI systems, strict actions must be taken against the liable institutions.[81]

It is a challenge to determine the accountability caused by the errors in AI systems. India must create a liability framework that assigns accountability to AI developers, deployers and the regulators.[82] Developers of AI systems should be legally responsible for creating bias-free and technically correct systems. Courts using AI technologies need to ensure that due process and rights are not infringed, and thus are liable for AI-driven legal injustices. Wrongfully convicted or unjustly ruled individuals, through the use of AI technology, should have the right to legal relief and compensation. A regime of strict liability, similar to that of the European Union's AI Liability Directive, can serve to ensure AI actors prioritise fairness and accountability in designing and implementing legal AI tools.[83]

DPDP AND GDPR: WHETHER USEFUL?

General Data Protection Regulation 2018 and Digital Personal Data Protection Act 2023 are the two significant legal frameworks intended to govern data collection, storage and processing in the European Union and India, respectively. Both the legislations are intended to harmonise

[79] Barocas and Selbst, ‘Big Data’s Disparate Impact’ (2016) 104 California Law Review 671
[80] AI Now Institute, ‘Algorithmic Impact Assessments’ (2018)
[81] European Commission, 'Liability Rules for Artificial Intelligence' (2022), European Parliament. https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en
[82] European Parliamentary Research Service, 'Artificial Intelligence Liability Directive' (2023), European Parliament
[83] Ibid

the technological progress and human rights, but their efficacy is questionable in addressing Artificial Intelligence-related challenges and the Black Box issue.

GDPR centres on individual rights and data transparency.[84] It governs both AI-driven and non-AI systems and mandates core principles like lawfulness, fairness, transparency, purpose limitation, data minimisation, and accountability. Article 5 mandates organizations to maintain transparency in data processing and demonstrate compliance with the GDPR principles. [85]

Articles 12-23 of GDPR provide the right to individuals to access, rectify, or erase their data if used unfairly. [86]

Article 22 of the GDPR limits automated decision-making without human intervention, which has direct effects on humans. It makes it possible for individuals to object to AI system decisions, which is important in legal domains such as predictive policing and sentencing prediction.[87]

GDPR promotes transparency in AI systems, but doesn't require AI systems to explain the decision-making logic of the systems. It also does not have any explicit regulation for algorithmic bias.

The DPDP Act 2023 was introduced to govern data privacy in India after the Supreme Court's decision in Justice K.S. Puttaswamy v. Union of India, declaring privacy a constitutional right. As opposed to GDPR, DPDP lays more emphasis on data governance than on the regulation of AI, hence it is less efficient in addressing the Black Box issue.[88] The legislation assures that digital personal data is treated in a manner that protects individuals' rights to safeguard their Personal Data and at the same time meets the necessity of processing such data for legal purposes and matters relating thereto.

The Act sets the groundwork for various other laws, such as the Digital India Act and industry-specific regulations concerning privacy and data protection. Section 4 of this Act requires explicit user consent for data collection. Section (6-11) ensures that the processing of personal

[84] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) [2016] OJ L 119/1, art 5.
[85] Ibid
[86] ibid
[87] ibid
[88] Justice K.S. Puttaswamy v Union of India (2017) 10 SCC 1.

data by the organisations must be reasonable and fair. Section 12 of this Act provides users the right to correct, update, complete, or erase their personal data.[89] However, there is no strong mechanism in this Act to audit AI algorithms for fairness or for law enforcement to disclose the decision-making process of systems like NATGRID or Facial recognition systems.

Aspect	**GDPR**	**DPDP**
Automated Decision-Making	Regulated under Article 22	No specific regulation
Right to Explanation	Granted under Articles 15 & 22	Not explicitly mentioned
Consent Mechanism	Explicit & Opt-In Required	Allows "Deemed Consent" for AI use
AI Bias & Fairness	General principles, but no strict regulation	No mention of AI fairness or bias
Surveillance & AI Policing	Limited government exceptions	Broad government exceptions for AI use

Both GDPR and DPDP are significant legislation against data protection and privacy, but they lack in regulating AI-driven decision-making systems. The lack of transparency, regulation of biases, and provision for AI accountability creates large gaps in both legislations, especially where AI is to be applied within the judiciary and law enforcement sectors. There is a pressing need for AI-specific laws in India to regulate AI within the legal system, requiring explainability, bias audits, and independent monitoring of AI-based judicial tools. In the absence of such reforms, AI-based legal decisions have the potential to undermine fundamental rights, fairness, and due process.

CONCLUSION

[89] Data Protection and Digital Privacy (DPDP) Act 2023

The inclusion of AI in the legal system promises efficiency but is accompanied by issues over bias, transparency, and accountability. India's reliance on generic laws such as the IT Act,[90] DPDP Act[91], and BNSS[92] makes AI-powered judicial tools unregulated, sparking doubts regarding due process and justice. AI-supported decisions lack explainability, putting fundamental rights at risk, and requiring a sound legal framework. Creating an AI regulatory body, mandating algorithm audits, and imposing strict liability regimes are necessary to provide accountability and redress for AI-based legal mistakes. AI must be kept as an aid, with judicial control providing equitable and just justice. Immediate legislative changes are required to bring AI applications in line with constitutional and ethical values in India's legal framework.

REFERENCES

- Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (Harvard University Press 2015).
- Yavar Bathaee, 'The Artificial Intelligence Black Box and the Failure of Intent and Causation' (2018) 31 Harvard JL & Tech 889.
- Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31 Harvard JL & Tech 841.
- European Union, Artificial Intelligence Act 2024 (EU AI Act 2024).
- Government of India, Digital Personal Data Protection Act 2023.
- National Crime Records Bureau, Automated Facial Recognition System: Implementation Report (2021).
- US Congress, Algorithmic Accountability Act 2023.
- White House Office of Science and Technology Policy, AI Bill of Rights 2022.
- Organisation for Economic Co-operation and Development (OECD), OECD AI Principles (2019).
- United Nations Educational, Scientific and Cultural Organisation (UNESCO), Recommendation on the Ethics of Artificial Intelligence (2021).
- Justice K S Puttaswamy (Retd) v Union of India (2017) 10 SCC 1.
- State v Loomis, 881 NW 2d 749 (Wis 2016).

[90] Information Technology Act, 2000
[91] Data Protection and Digital Privacy (DPDP) Act, 2023.
[92] Bhartiya Nagrik Suraksha Sanhita (BNSS), 2023.

- R (Bridges) v South Wales Police [2020] EWCA Civ 1058.
- https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- https://www.technologyreview.com/2021/01/07/1016221/whats-wrong-with-ai-crime-prediction-everything/
- https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html/
- https://bigbrotherwatch.org.uk/
- https://internetfreedom.in/automated-facial-recognition-report/
- https://www.spiceworks.com/tech/artificial-intelligence/articles/ai-regulations-around-the-world/