
BEYOND THE BLACK BOX AND THE DUAL CONSTITUTIONAL CRISIS OF HEALTHCARE AI IN INDIA

Mohamed Asraf¹ & Reema²

ABSTRACT

Clinical medicine is undergoing a quiet but consequential transition: rule-based clinical software, whose logic could be audited line by line, is giving way to generative and deep-learning CDSS whose reasoning even their developers cannot fully reconstruct. This shift produces a dual crisis for Indian health jurisprudence. At the systemic level, it tests the positive obligation to furnish equitable healthcare that the Supreme Court read into Article 21 of the Constitution. At the individual level, it strains the doctrine of informed consent, since a patient cannot meaningfully authorize a diagnostic pathway that no one in the room can explain. This paper argues that Indian law currently addresses these two dimensions in isolation — constitutional health jurisprudence through equality and access-based reasoning, and consent through common-law disclosure standards — without a unifying account of what algorithmic opacity does to both simultaneously. Reading Article 21 and Article 14 jurisprudence alongside the disclosure standard set out in *Samira Kohli v. Dr. Prabha Manchanda* (2008) and the fiduciary obligations created by the Digital Personal Data Protection Act, 2023, the paper develops legal standards for algorithmic explainability, non-discriminatory health data governance, and statutory consent mechanisms suited to black-box clinical AI, and proposes a stratified consent model and a modified negligence standard to operationalize them.

Keywords: algorithmic opacity, right to health, informed consent, Digital Personal Data Protection Act 2023, explainable AI, medical negligence.

¹ Assistant Professor, Saraswathy Law College, Tindivanam, affiliated under The Tamilnadu Dr. Ambedkar Law University, Chennai.

² Assistant Professor, Mother Teresa Law College, Illupur, affiliated under The Tamilnadu Dr. Ambedkar Law University, Chennai.

Introduction

Clinical software has moved through three broad generations. The first generation was rule-based expert systems built on explicit “if-then” logic that a physician, or a court, could trace step by step. The second added statistical machine learning, still largely interpretable through documented feature weights. The current generation, built on deep neural networks and increasingly on generative foundation models, trades this interpretability for diagnostic performance. A Clinical Decision Support System (hereafter referred to as CDSS) may flag a lesion as malignant with high confidence while offering no auditable account of which pixels, patterns, or correlations drove that judgment. This is the “black-box” problem, and it is no longer confined to research settings; it is entering Indian tertiary hospitals through radiology, pathology, and triage tools procured from both domestic and multinational vendors.

The transition produces what this paper calls a dual crisis. The first is systemic: Article 21’s guarantee of accessible, quality healthcare was built for a world of scarce doctors and scarce beds; it did not anticipate a world where the quality of care itself might depend on a diagnostic model trained on data that under-represents the population being treated. The second is individual: the doctrine of informed consent, refined in Indian tort and consumer law around a physician’s capacity to disclose material risks, presumes that someone can articulate what the risk is? If Black-box diagnosis unsettles that premise directly. Against this backdrop, the paper pursues three linked objectives: first, to establish what standard of algorithmic explainability Article 21 and Article 14 jurisprudence should be read to require of clinical AI; second, to determine what a legally valid consent process looks like when the treating physician cannot fully explain the system’s internal logic; and third, to map how the Digital Personal Data Protection Act, 2023 structures the statutory consent and fiduciary obligations that apply once patient data enters such a system.

AI, Algorithmic Bias, and the Constitutional Right to Health

The starting point for any constitutional inquiry into AI-driven diagnosis is the line of Article 21 jurisprudence that reads the right to life as encompassing a right to health. In *Paschim Banga Khet Mazdoor Samity v. State of West Bengal*,³ the Supreme Court held that a state hospital’s refusal to admit an accident victim for want of a bed violated Article 21, and that the State’s

³*Paschim Banga Khet Mazdoor Samity v. State of West Bengal*, AIR 1996 SC 2426.

financial constraints could not excuse a failure to provide timely, adequate medical treatment. The judgment is significant less for its facts than for the structure of obligation it imposes, it converts healthcare from a policy aspiration into a justiciable entitlement, and it does so specifically at the point of access were the moment a patient meets, or fails to meet, the health system's capacity to treat them.

Algorithmic diagnosis relocates that point of failure. A hospital that has beds, doctors, and equipment can still fail a patient if the CDSS interpreting their scan or their vitals was trained on a population that does not resemble them. This is not a hypothetical concern. A recent multidimensional review of AI-driven diagnostics found that underrepresentation of rural populations in training datasets was linked to a substantially elevated false-negative rate for a common respiratory diagnosis, while dermatological AI systems showed markedly higher error rates for patients with darker skin due to imbalanced training data.⁴ At a global level, a widely cited review of clinical AI literature found that more than half of the datasets used to build and validate such systems originate from the United States or China, with over 40% of authorship also concentrated in those two countries explicit that most clinical AI is built for, and validated on, populations that bear little resemblance to rural or peri-urban India.⁵ Where an Indian hospital deploys such a system without local validation, the “failure to provide adequate treatment” that Paschim Banga condemns may occur silently, inside the model, well before any doctor or administrator is in a position to notice it.

This misclassification risk is also an equality problem, not merely an access problem. Article 14's guarantee against arbitrary and unequal treatment has long been read, alongside Article 21 to prohibit state action and, through the horizontal reach of public health regulation, private action delivering a public function, produces systematically worse outcomes for a class of persons without a rational and proportionate justification. An AI diagnostic tool that is statistically more accurate for urban, English-literate, higher-income patients than for rural, low-literacy, lower-income patients does not discriminate on its face; it discriminates through its error distribution. Indian equality jurisprudence has not yet had occasion to test this species of indirect, statistically mediated discrimination against a clinical algorithm, but the doctrinal

⁴Li, Y., Yi, X., Fu, J., Yang, Y., Duan, C., & Wang, J. (2025). Reducing misdiagnosis in AI-driven medical diagnostics: A multidimensional framework for technical, ethical, and policy solutions. *Frontiers in Medicine*, 12, 1594450. <https://doi.org/10.3389/fmed.2025.1594450>

⁵Celi, L. A., Cellini, J., Charpignon, M.-L., Dee, E. C., Dernoncourt, F., et al. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health*, 1(3), e0000022. <https://doi.org/10.1371/journal.pdig.0000022>

materials disparate impact reasoning under Article 14 combined with the substantive content of the Article 21 health right are already available to do so.

The equity problem compounds when AI diagnostics are treated as instruments of public health scaling rather than public health equalization. The same infrastructural gap that has historically kept rural India underserved by specialists now threatens to determine who benefits from AI-assisted diagnosis at all to high-performance CDSS tools are frequently deployed through proprietary, cloud-hosted, subscription-priced platforms concentrated in urban tertiary and private hospitals, while public health centres were precisely the facilities situation of Paschim Banga obliges the State to adequately staff and equip are lacks the bandwidth, hardware, and licensing budgets to deploy the same tools. Without a deliberate policy choice favouring open, publicly funded diagnostic AI infrastructure for primary and secondary care, algorithmic medicine risks becoming a second digital divide layered directly on top of India's existing rural-urban health access divide improving outcomes precisely where outcomes are already comparatively good, and leaving the constitutional promise of Paschim Banga unfulfilled exactly where it matters most.

The Crisis of Informed Consent in Black-Box Systems

Indian informed-consent doctrine reached its modern form in *Samira Kohli v. Dr. Prabha Manchanda*,⁶ where the Supreme Court rejected the American “reasonable patient” standard in favour of a version of the English Bolam-linked standard have doctor must disclose the nature and procedure of the treatment, its material risks, and any available alternatives, in terms a patient of ordinary intelligence can understand, before undertaking any procedure beyond what was consented to. The Court was emphatic that consent obtained for a diagnostic procedure does not extend to a therapeutic one performed in the same sitting, even where the additional procedure is, from a purely clinical standpoint, in the patient's best interest. The doctrine's entire architecture presupposes a narratable risk: the physician must be able to say, in advance, what might go wrong and why.

Black-box CDSS breaks this architecture at its foundation. Where a rule-based system flags a risk, the rule itself is the explanation as “elevated troponin plus this ECG pattern indicates X” is both the output and the disclosure. A deep neural network trained end-to-end on imaging or

⁶Samira Kohli v. Dr. Prabha Manchanda, (2008) 2 SCC 1.

genomic data instead produces a probability score derived from millions of weighted parameters interacting in ways that resist human-legible summary; post-hoc explainability tools (saliency maps, SHAP values, and similar techniques) offer approximations of what influenced the model, not a reconstruction of its actual reasoning. When a physician tells a patient that “the AI system rates this scan as 85% likely to be malignant” without being able to say why, the disclosure satisfies the letter of “informing the patient of the nature of the procedure” while hollowing out its substance. Samira Kohli requires disclosure of material information the physician possesses; it does not obviously contemplate a scenario where the physician relies on a determination that neither they nor, in a strict sense, anyone else fully possesses.

This raises a genuine doctrinal question rather than a merely rhetorical one, can consent to AI-assisted diagnosis ever be “informed” in the Samira Kohli sense, or does the doctrine need to be reconstructed around a different object of disclosure are not the AI’s internal logic, but its known error characteristics, validation population, and confidence calibration? The more defensible reading is the latter. Just as a patient consenting to a genetic test is not owed an explanation of the biochemistry of PCR amplification, a patient consenting to AI-assisted diagnosis may not need the model’s internal logic explained, but they are owed disclosure of what the tool has and has not been validated to do crucially, whether it has been validated on a population resembling them. Read this way, the crisis of consent in black-box systems is not fundamentally a technical-explainability problem but a disclosure-scope problem: existing consent forms and practices in Indian hospitals were not built to capture this category of information, and nothing in the Samira Kohli framework currently obliges through a baseline level, standard medical law strictly requires explicit, informed consent before any biological sample (such as a PCR swab or blood draw) is collected, with exceptions strictly limited to life-threatening emergencies, proxy guardian consent for incapacitated individuals, or statutory criminal mandates. When that physical sample is transformed into digital health data, the downstream digital processing demands an equally rigorous standard: feeding patient diagnostic data into a CDSS is permissible for direct clinical decision-making, but secondary algorithmic training or commercial analytics necessitates independent, unbundled consent."

A second and distinct problem compounds the first: automation bias, the well-documented tendency of decision-makers to over-rely on automated recommendations and under-scrutinize

them, including when those recommendations are wrong.⁷ A systematic review of the clinical decision-support literature found that automated advice reliably shifts clinician judgment, at times causing clinicians to override an initially correct assessment in favour of an incorrect algorithmic one, with the effect more pronounced under time pressure and heavier caseloads.⁸ Conditions that describe the majority of Indian public-hospital outpatient departments. Automation bias matters to consent doctrine because Samira Kohli's framework assumes an exercising physician: someone actively weighing the treatment decision, of which the patient's consent is meant to be a genuine check. Where automation bias causes the physician to defer near-automatically to the CDSS output, the physician's independent clinical judgment varies because of fiduciary relationship, and the disclosure duty flowing from it, which is designed to protect or displace. The consent process risks becoming a formality wrapped around a decision the physician did not, in any meaningful sense, independently make. This is not a hypothetical concern in a jurisdiction where consumer-protection litigation already treats a doctor's failure to exercise independent judgment as a component of deficiency in service; it is a foreseeable extension of existing negligence and consumer-law reasoning to a new class of instrument.

Data Governance, Privacy, and Data Fiduciary Duties

Every AI-assisted diagnosis begins with data leaving the patient's body and entering a computational pipeline, and Indian law now regulates that movement through two overlapping statutory regimes: the Digital Personal Data Protection Act, 2023 ("DPDP Act"), and the residual provisions of the Information Technology Act, 2000 ("IT Act"). Understanding how they interact is a precondition for any workable consent architecture for clinical AI.

Under the IT Act, Section 43A historically imposed civil liability on a "body corporate" which includes hospitals and diagnostic chains for failing to implement "reasonable security practices" in handling sensitive personal data, including medical information, backed by the 2011 Sensitive Personal Data Rules.⁹ Section 44(2) of the DPDP Act formally omits Section 43A,¹⁰ but the transition is being phased rather than immediate industry guidance indicates the IT Act regime and the 2011 Rules will continue to operate in parallel until the DPDP Act's

⁷Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127. <https://doi.org/10.1136/amiainl-2011-000089>

⁸Id.

⁹Information Technology Act, 2000, § 43A (India); Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011 (India).

¹⁰Digital Personal Data Protection Act, 2023, § 44(2)(a) (India).

own security and breach-notification provisions are fully notified, with a transition currently expected around 2027.¹¹ For the present, then, a hospital or AI vendor processing clinical data sits under a dual security-compliance regime. Section 72A of the IT Act remains independently in force and is not affected by this transition: it criminalizes the unauthorized disclosure of personal information obtained under a lawful contract, with imprisonment up to three years or a fine up to five lakh rupees provision directly relevant where a hospital's data-sharing arrangement with a third-party AI vendor is breached or exceeded.¹²

The DPDP Act supplies the newer and more comprehensive framework. It designates any entity that determines the purpose and means of processing personal data squarely including a hospital deploying diagnostic AI, as a "Data Fiduciary," and imposes on it a notice-and-consent architecture built around Sections 5 and 6. Section 5 requires that, before or alongside any request for consent, the Data Fiduciary gives the patient a notice describing the personal data to be processed, the specific purpose of that processing, and how the patient may exercise their rights or lodge a complaint.¹³ Section 6 defines valid consent as free, specific, informed, unconditional, and unambiguous, expressed through clear affirmative action, and confined to the purpose stated in the notice with the burden resting on the Data Fiduciary to prove, if the matter is ever disputed, that the required notice was given and that consent was validly obtained.¹⁴ Read together with the DPDP Rules, 2025, the notice must be presented independently of other documents, in plain language, and must itemize the specific data categories and purposes involved¹⁵ a standard that a bundled, one-line consent form authorizing "use of your data for diagnosis and treatment" is unlikely to satisfy once an AI system, and potentially a third-party AI vendor, is added to that processing chain. In practice, this establishes a clear dividing line between legitimate clinical utility and unlawful processing: integrating laboratory findings into an internal CDSS solely to alert the treating physician about immediate diagnostic risks aligns with purpose limitation principles under data protection laws. Conversely, transmitting identifiable patient metrics to third-party proprietary vendors or burying digital tracking permissions within multi-page admission forms invalidates the statutory consent architecture.

¹¹Chambers and Partners. (2026). India's Digital Personal Data Protection Regime Takes Effect. <https://chambers.com/articles/s-r-data-india-s-digital-personal-data-protection-regime-takes-effect>

¹²Information Technology Act, 2000, § 72A (India).

¹³Digital Personal Data Protection Act, 2023, § 5 (India).

¹⁴Digital Personal Data Protection Act, 2023, § 6(10) (India).

¹⁵Digital Personal Data Protection Rules, 2025, Rule 3 (India).

This creates a structural gap that Indian hospitals have generally not yet closed. Where clinical data is transmitted to a third-party proprietary AI platform frequently cloud-hosted, sometimes outside India for diagnostic inference, the hospital's Section 5 notice must, to be DPDP-compliant, disclose that transmission as a distinct processing purpose, name or meaningfully describe the recipient, and allow the patient to understand that a non-human, non-clinician system will process their data. An open-source or locally hosted model raises a narrower but still live version of the same problem: even where data does not leave institutional custody, the purpose of "AI-assisted diagnostic processing" is analytically distinct from "clinician review" and, under Section 5's itemization requirement, arguably needs to be separately disclosed and separately consented to, not folded into a generic treatment-consent form. A Data Fiduciary that processes health data through an AI system without disclosing that fact as a specific purpose is, on a plain reading of Sections 5 and 6, collecting consent that does not cover the processing actually occurring and rendering that consent legally infirm under the DPDP Act quite independently of whether it also satisfies (or fails) the Samira Kohli standard discussed above. The two frameworks, constitutional/tort-based medical consent and statutory data-fiduciary consent, converge on the same practical requirement without, at present, being read together in Indian hospital compliance practice.

Normative Solutions and Regulatory Framework

The preceding sections show that Indian law currently regulates AI-driven diagnosis through three unconnected channels such as constitutional health jurisprudence, common-law informed consent, and statutory data-fiduciary obligations each adequate to a world of legible, rule-based software and each strained by opacity. Three normative reforms would connect them.

First, mandate explainability as a precondition of regulatory approval, not a voluntary best practice. Encouragingly, Indian regulatory practice has already begun moving in this direction. The CDSCO's final Guidance Document on Medical Device Software (July 2026) requires applicants for AI-enabled medical device software to disclose how representative their training, validation, and testing data are of the population the tool is meant to serve, together with evidence of how the model's error rates and reliability vary across relevant patient sub-groups, and whether that data was drawn from real clinical records or generated synthetically.¹⁶ This is

¹⁶Central Drugs Standard Control Organisation, Guidance Document on Medical Device Software, Doc. No. CDSCO/MD/GD/MDSW/01/2026 (July 21, 2026).

a genuine advance, but it remains a pre-market disclosure obligation directed at the regulator rather than a point-of-care explainability obligation directed at the treating physician or patient. The Guidance should be extended, through a CDSCO circular or amendment to the Medical Devices Rules, 2017, to require that every AI diagnostic tool cleared for clinical use ship with a standardized, plain-language “model characteristics statement” population validated on, known error rates by relevant sub-group, and confidence-calibration information that hospitals are obliged to translate into the Section 5 DPDP notice and the Samira Kohli consent conversation. Explainability, in other words, should be operationalized not as demanding a technically impossible account of a neural network’s internal weights, but as guaranteeing the disclosure-scope reconstruction proposed in Section 3.

Second, adopt a stratified, two-tier consent model. Administrative AI have scheduling, billing, record transcription, triage queue management leads to carries negligible clinical risk and can reasonably continue under general-purpose data-processing consent, provided the DPDP Act’s Section 5 notice requirements are met. Autonomous diagnostic or interventional AI systems that generate a diagnosis, risk score, or treatment recommendation materially relied upon in clinical decision-making, should require a distinct, itemized consent instrument, executed separately from general treatment consent, disclosing: (i) that an AI system participated in the diagnostic process; (ii) the model characteristics statement referred to above; (iii) whether a qualified clinician independently reviewed the AI output or substantially deferred to it; and (iv) the patient’s right to request purely clinician-based diagnosis where clinically feasible and (v) an explicit right to opt out of secondary software telemetry and algorithmic retraining while retaining full access to standard, clinician-led medical treatment. This mirrors, in structure, the separate-consent requirement Samira Kohli itself imposed for procedures exceeding the scope of initial consent has to extend a principle already native to Indian medical jurisprudence to a new category of clinical instrument, rather than importing a foreign doctrinal transplant.

Third, adapt the medical negligence standard to apportion algorithmic-diagnostic-failure liability. *Jacob Mathew v. State of Punjab*¹⁷ set a demanding threshold for criminal medical negligence requiring “gross” negligence reflecting a marked departure from the standard of a reasonably competent practitioner, while leaving civil negligence to the ordinary Bolam-derived duty-of-care standard. Neither limb currently distinguishes between a clinician’s own diagnostic error and an error the clinician imported by relying on a flawed AI output. A

¹⁷*Jacob Mathew v. State of Punjab*, (2005) 6 SCC 1.

workable apportionment framework would ask, sequentially: (a) did the AI tool operate within its disclosed validation parameters (a question resolved largely by the CDSCO dataset-disclosure record proposed above); (b) if it did not for instance, misclassifying a patient from a demographic the tool was never validated on liability should presumptively shift toward the manufacturer and the institution that deployed an inadequately validated tool without disclosure, rather than the individual clinician; and (c) if the tool operated within disclosed parameters but the clinician exhibited automation bias by disregarding contrary clinical signs, ordinary Bolam-based clinician liability should continue to apply, informed by Jacob Mathew's gross-negligence threshold for any criminal charge. This tiered approach avoids both extremes current practice risks drifting toward: absolving clinicians entirely because "the algorithm decided," and holding clinicians to an impossible standard of independently verifying opaque model outputs they were never equipped to audit.

Conclusion

Algorithmic opacity in clinical diagnosis is not two separate problems awaiting two separate doctrinal answers. It is a single structural challenge to the promise that Article 21 makes and the promise that informed consent is supposed to keep. Paschim Banga obliges the State to make quality healthcare accessible; black-box AI risks making that quality silently uneven along exactly the lines of rural-urban and demographic disadvantage the Constitution guards against. Samira Kohli obliges the physician to disclose what they know; black-box AI produces outputs that resist being fully known, requiring consent doctrine to pivot from explaining how a system reasons to disclosing what it has been validated to do. The Ministry of Health & Family Welfare and CDSCO are well placed to close this gap without new primary legislation: by extending the CDSCO's existing bias-disclosure requirements from a pre-market filing into a point-of-care disclosure standard, by mandating the two-tier consent model proposed here for any AI system materially informing diagnosis, and by issuing interpretive guidance apportioning negligence liability between institution, manufacturer, and clinician along the lines set out above. Absent such intervention, India risks importing the efficiency gains of clinical AI while quietly exporting its costs to the patients least able to contest them.

References

1. Celi, L. A., Cellini, J., Charpignon, M.-L., Dee, E. C., Dernoncourt, F., et al. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health*, 1(3), e0000022. <https://doi.org/10.1371/journal.pdig.0000022>
2. Central Drugs Standard Control Organisation. (2026). Guidance Document on Medical Device Software (Doc. No. CDSCO/MD/GD/MDSW/01/2026).
3. Chambers and Partners. (2026). India's Digital Personal Data Protection Regime Takes Effect. <https://chambers.com/articles/s-r-data-india-s-digital-personal-data-protection-regime-takes-effect>
4. Digital Personal Data Protection Act, No. 22 of 2023, INDIA CODE (2023).
5. Digital Personal Data Protection Rules, 2025 (India).
6. Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121–127. <https://doi.org/10.1136/amiajnl-2011-000089>
7. Information Technology Act, No. 21 of 2000, INDIA CODE (2000).
8. Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011 (India).
9. Jacob Mathew v. State of Punjab, (2005) 6 SCC 1 (India).
10. Li, Y., Yi, X., Fu, J., Yang, Y., Duan, C., & Wang, J. (2025). Reducing misdiagnosis in AI-driven medical diagnostics: A multidimensional framework for technical, ethical, and policy solutions. *Frontiers in Medicine*, 12, 1594450. <https://doi.org/10.3389/fmed.2025.1594450>
11. Paschim Banga Khet Mazdoor Samity v. State of West Bengal, AIR 1996 SC 2426 (India).
12. Samira Kohli v. Dr. Prabha Manchanda, (2008) 2 SCC 1 (India).