International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

AI as a Decision-Maker. Where Should Human Judgment Still Be Mandatory?

Mr. Abhay Chauhan[1], Dr. Simarjeet Kaur[2]

[1]Student, CSE

Abstract

A decade ago, artificial intelligence was something you read about in research journals or saw demoed at a conference. Today it sits inside everyday work. Doctors use it to flag anomalies on scans. Judges in some jurisdictions get risk scores before they set bail. Lenders run applications through credit models before a human ever opens the file. Even teachers, in classrooms that use adaptive platforms, are getting nudges about which student to check on. The question this paper takes up is narrow but pressing: when do we let the machine decide, and when do we keep a human in the chair?

The study is built on secondary sources — peer-reviewed research, regulatory texts, industry reports, and well-documented case studies — drawn mostly from 2010 through 2024. The approach is qualitative and descriptive. Numbers alone don't resolve questions about who should be accountable when a decision goes wrong, and the literature on procedural fairness has been making this point for years.

AI is excellent at some things and poor at others. It handles volume, structure, and pattern recognition well. It struggles when a decision turns on context, values, or the kind of accountability that needs a person who can answer for it. The argument here is not that AI is dangerous, nor that it should be embraced without reservation. It is that decision systems work best when they are hybrid by design — machines doing what machines do well, humans owning what humans uniquely can. A three-tier classification proposed in Section 9 is the practical contribution. But the framework only matters if the institutions around it — hospitals, courts, banks, schools, regulators — are willing to rethink how they delegate authority and assign blame when something breaks.

Keywords: artificial intelligence, decision-making, human judgment, algorithmic bias, AI ethics, human–AI collaboration, responsible AI, hybrid oversight model

1. Introduction

Is data really driving the big decisions now? A radiologist I spoke to in Chennai described the software on his workstation in almost casual terms. It highlights anomalies, suggests where to look closer, helps him decide whether a patient needs more imaging. He wasn't anxious about it. He also wasn't dismissive. It was just part of the job, the way a stethoscope or a referral pad is. A loan officer in Mumbai told a similar story, and so did a recruiter in Bengaluru — algorithms surface candidates or risk scores, and the human stamps the final call. We aren't fully automated. But the direction is hard to argue with.

The case for algorithmic help is genuine. Machines don't get tired or distracted. They can hold more variables in mind at once than any person could. They don't have a bad day before lunch. That matters

in fraud detection, in radiology reads, in real-time traffic systems. Kahneman's (2011) "System 1" — the fast, intuitive pattern-matching that humans rely on for snap judgments
— is exactly what machines do natively. "System 2," the slow and deliberative kind, is where their record gets shakier.

Most consequential decisions, though, aren't pattern-matching problems. When a judge sentences someone, she has to weigh remorse, family circumstances, the odds of rehabilitation, and the social cost of incarceration. When a doctor sits with a patient who is unlikely to recover, the conversation turns on what the patient values, what the family can carry, and how to talk about dying without lying. When a teacher tries to read a quiet student, the question is whether the student is bored, sad, struggling, or simply not seen — and the response has to fit the actual reason. Philosophers have a name for this kind of work. They call it practical wisdom. No one has yet figured out how to write it as code.

This paper is not in the business of claiming AI is bad and humans are good. The behavioural economics literature has already made the opposite case persuasively (Tversky and Kahneman, 1974). The harder question is design: how do you build a decision system that uses both kinds of intelligence in the right proportion, so that what comes out the other end is accurate, fair, and defensible if someone challenges it?

The paper is organised as follows. Section 2 reviews the literature. Section 3 sets out the method. Sections 4 through 8 examine five domains where the question is live — healthcare, finance, criminal justice, recruitment, and education. Section 9 pulls the cross-domain pattern together and proposes a three-tier taxonomy. Section 10 looks at the wider ethical stakes. Section 11 closes with recommendations and the open questions that remain.

2. Literature Review

2.1 The Rise of Algorithmic Decision-Making

Algorithmic decision-making is older than the current AI conversation suggests. The expert systems of the 1970s and 1980s already automated parts of professional judgment — they just did it through hand-coded rules. The shift that matters today is the move from explicit rules to learned ones. Modern deep networks derive their own decision logic from training data, and that logic is often inscrutable even to the engineers who built the model (Doshi-Velez and Kim, 2017).

This is the "black box" problem. Burrell (2016) breaks it down into three kinds of opacity: deliberate (the model is a trade secret), technical (most users don't have the expertise to interrogate it), and intrinsic (the math itself doesn't produce clean human-readable explanations). Each kind calls for a different fix. What they share is that algorithmic decisions become genuinely hard to challenge once they're made.

The opacity problem runs parallel to a second one: documented bias in deployed systems. ProPublica's 2016 investigation into COMPAS, a recidivism prediction tool used across US courts, is the most cited example. The analysis showed the system was almost twice as likely to misclassify Black defendants as high-risk than similarly situated white defendants (Angwin et al., 2016). The reporting set off a long debate about what "fairness" even means mathematically
— a debate that hasn't fully resolved, partly because Chouldechova (2017) showed that several reasonable-sounding definitions are mathematically incompatible with each other.

2.2 Human Judgment: Capabilities and Systematic Failures

It would be a mistake to treat human judgment as the unproblematic alternative to algorithmic decision-making. Decades of behavioral economics research have mapped the systematic ways in which people's judgments deviate from rational norms. Anchoring effects, availability bias, confirmation bias, and in-group favoritism are among the most consistently replicated findings in social psychology. People rely on cognitive shortcuts — heuristics — that serve them reasonably well most of the time but can yield seriously flawed conclusions under pressure, particularly when processing complex information quickly (Tversky and Kahneman, 1974).

The empirical record bears this out in decision-relevant contexts. Dressel and Farid (2018) found that untrained humans asked to assess recidivism risk performed no better than COMPAS on average — and displayed similar racial disparities. This finding shifts the terrain of the debate considerably. The question is less "AI or human?" and more "which arrangement of human and machine judgment, structured in what way, produces the most defensible outcomes?"

What human decision-making may offer that current AI cannot — at least not reliably — is moral agency, contextual sensitivity, and the structural capacity for accountability. When a person makes a decision, there is, at least in principle, someone who can be held responsible, who can articulate a rationale, and who bears some felt obligation toward the consequences. These properties matter independently of predictive accuracy. Procedural justice research suggests that people's perceptions of fairness are shaped as much by how decisions are made as by their substantive outcomes (Tyler, 2003).

2.3 Human-AI Collaboration Models

A third strand of the literature moves past the either/or framing and asks how the two can be combined productively. Wilson and Daugherty (2018) call it "amplification" — AI surfaces patterns, flags inconsistencies, broadens the option set, and a human interprets and decides. In clinical practice, the "clinician-in-the-loop" model has become a regulatory expectation in most major markets: an AI recommendation has to be reviewed and signed off by a licensed practitioner before action is taken (Topol, 2019). In criminal justice, Stevenson and Doleac (2019) have argued for structured review protocols that keep AI inside a wider frame of human accountability, rather than letting it operate alongside or beneath the human.

Across all three strands, the literature broadly points in one direction, which this paper builds on: the right amount of human involvement in an AI-assisted decision should track the stakes of the decision, how reversible it is, and how much contextual or moral judgment it actually requires.

3. Methodology

3.1 Research Design

The design is qualitative and descriptive, built around systematic secondary-data analysis. The choice fits the questions being asked. Working out where human oversight should be mandatory requires interpretive reasoning, not just measurement. Concepts like moral accountability, social legitimacy, and procedural fairness resist clean operationalisation, and the method has to match.

3.2 Data Sources

Four kinds of sources fed the study. Peer-reviewed work was pulled from Google Scholar, JSTOR, IEEE Xplore, and the ACM Digital Library, using search terms like "AI decision-making," "algorithmic bias," "human-AI collaboration," and sector-specific variants. The review favoured work from 2010 onward, and especially from 2018, the point at which deep learning systems started being deployed at meaningful

institutional scale.

Industry reports from McKinsey Global Institute, Deloitte, the World Economic Forum, and the Partnership on AI were read alongside the academic material. Regulatory and government documents — the EU AI Act (European Commission, 2021), the OECD AI Principles (OECD, 2019), India's National Strategy for Artificial Intelligence (NITI Aayog, 2018), and the 2023 US Executive Order on AI — were analysed to capture the governance landscape. Documented deployments and failures, drawn from investigative journalism, legal proceedings, and academic case studies, rounded out the source base.

3.3 Analytical Approach

The approach was thematic. Material was coded around three primary themes: what AI is actually doing well or poorly in each domain, what human judgment specifically adds, and what governance mechanisms have been built or proposed. A secondary pass picked up the recurring ethical issues — bias, opacity, accountability, autonomy — and the conditions under which human oversight appears genuinely necessary rather than just precautionary.

Sections 4 through 8 work through the domains one at a time. Section 9 then steps back to look at what they have in common. The three-tier taxonomy proposed in that section is the study's original contribution. It came out of the coding work, and it is meant to be usable by the people who actually have to decide where AI belongs and where it doesn't.

4. AI Decision-Making in Healthcare

4.1 Current Applications

Healthcare is probably the domain where AI's upside and its risks are both clearest. Current deployments span medical imaging, clinical decision support, drug discovery, triage, and back-office work. IBM Watson for Oncology was one of the first high-profile attempts at clinical recommendation, and its arc is instructive. When internal documents from Memorial Sloan Kettering's partnership with Watson became public, they showed the system had produced treatment suggestions described internally as "unsafe and incorrect" in several cancer cases (Ross and Swetlitz, 2018).

When the scope is narrower, the record is better. DeepMind's 2018 retinal-disease work showed that a deep-learning system could identify a wide range of retinal pathologies from scans with accuracy comparable to specialist ophthalmologists (De Fauw et al., 2018). Comparable results have been reported in dermatology, radiology, and pathology — fields where the core task is exactly what deep learning does best: pattern recognition on high-volume visual data. In India, AI-assisted screening for diabetic retinopathy has been piloted in several states, and the early results suggest that even general practitioners in primary health centres can use such tools to catch cases that would otherwise be referred too late.

4.2 Where Human Judgment Remains Irreplaceable

What the successful healthcare AI applications have in common is a tightly scoped task: identifying patterns in structured data. The rest of the clinical workflow — taking a history, understanding the patient's values, navigating the family, explaining uncertainty without making things worse, and choosing a path that fits the whole person — stays irreducibly human, at least for now.

Informed consent is the cleanest illustration. Consent isn't a checkbox. It is the process of helping a patient understand what's happening, what their options are, what's still uncertain, and then supporting

their right to choose. It is relational and ethical work. You cannot outsource it to a system that has no model of what illness feels like, what this particular patient cares about, or what their family can absorb. Obermeyer and Emanuel (2016) make a related point: the hardest part of medicine is often not figuring out what's wrong but deciding what to do once you know.

These limits aren't merely theoretical. Obermeyer et al. (2019) studied a widely deployed US algorithm used to identify patients in need of high-risk care management, and found it systematically under-estimated the clinical needs of Black patients. The mechanism is worth understanding because it shows up everywhere. The algorithm didn't use race as a variable. It used historical healthcare spending as a proxy for need. But spending under-represents Black patients' actual needs, because Black patients historically had less access to care. To catch a bias like that, you need someone scrutinising the social assumptions baked into the proxy — and the system can't do that for itself. Mplications The regulatory consensus that has emerged — reflected in the EU's Medical Device Regulation, the US FDA's framework for AI/ML-based software, and India's developing regulatory landscape — is that clinical AI must function as a decision support tool, not an autonomous decision-maker. Final clinical authority is expected to remain with licensed, accountable human practitioners. This is not simply conservative caution. It reflects a substantive judgment that accountability and the relational dimensions of clinical care cannot be structurally separated from technical competence in medicine without real harm to patients and to the legitimacy of healthcare institutions themselves.

4.3 Healthcare Governance

The regulatory consensus that has emerged — visible in the EU's Medical Device Regulation, the US FDA's framework for AI/ML-based software, and the developing rules in India — is consistent. Clinical AI is decision support, not an autonomous decision-maker. Final authority stays with the licensed practitioner. This is not just risk aversion. It reflects a substantive judgment that the accountability and relational dimensions of medicine cannot be detached from the technical work of diagnosis without doing real damage — both to patients and to the legitimacy of the institutions involved.

5. AI Decision-Making in Finance

5.1 Applications and Performance

Finance has gone the furthest down the automation road, and on practical grounds it had the easiest case to make. Credit scoring, fraud detection, algorithmic trading, and risk modelling all involve large, well-structured datasets, clear optimisation targets, and tight feedback loops. FICO scores in the US — and the CIBIL score in India — are themselves a kind of algorithmic scoring, and they predate the current AI conversation by decades.

Modern machine learning extends the toolkit considerably. Models trained on behavioural and alternative data can include signals that traditional scoring couldn't process. In fraud detection, major banks now run AI across hundreds of millions of daily transactions, surfacing suspicious patterns in real time at a scale no human team could match. JPMorgan Chase's Contract Intelligence platform reportedly processed 12,000 commercial credit agreements in seconds — work previously estimated at 360,000 hours of legal review annually (Corkery and Protess, 2017). In India, the Unified Payments Interface has produced a scale of transaction data that has made AI-based fraud monitoring not optional but essential.

5.2 Embedded Bias and Fairness Concerns

The performance is real. So are the failures. The 2019 Apple Card episode — in which an algorithm-driven credit system assigned very different limits to married couples with apparently similar financial profiles on the basis of gender — drew a regulatory investigation from the New York Department of Financial Services. Amazon's discontinued AI hiring tool, which was found to be downgrading resumes that referenced women's clubs and colleges, surfaced the same underlying issue from a different angle. A model trained on past decisions carries forward whatever biases those decisions contained.

These weren't implementation errors. They reflect a structural property of supervised learning. Training data shaped by decades of redlining, discriminatory underwriting, and uneven access will produce models that reproduce those patterns unless someone steps in deliberately. That intervention requires social and historical judgment — about which variables are legitimate signals of risk and which are just historical discrimination dressed up in technical clothing. The system cannot make that distinction on its own.

5.3 The Systemic Risk Dimension

Financial markets also raise a concern that's largely absent in other domains. When many participants run similar models on similar data, they may respond to the same signals in correlated ways, and amplify volatility rather than dampen it. The May 2010 Flash Crash, in which the Dow Jones Industrial Average dropped roughly 1,000 points in minutes before partially recovering, was attributed in part to self-reinforcing behaviour across automated trading systems. The implication is that financial oversight cannot only be about individual decisions. It has to extend to the architecture of the market and to the aggregate behaviour of the systems running inside it.

6. AI Decision-Making in Law Enforcement and Criminal Justice

6.1 Predictive Policing and Recidivism Assessment

Criminal justice is where the tension is sharpest. The efficiency upside is genuine. The human-rights cost, when the system gets it wrong, is severe. Predictive policing tools — which forecast where crime is likely or who is likely to offend — have been deployed across the United States, the United Kingdom, and several other jurisdictions. Recidivism tools like COMPAS have shaped bail, sentencing, and parole decisions for hundreds of thousands of people.

The bias evidence here is substantial. Predictive policing models trained on historical arrest data inherit the demographic and geographic skew of past enforcement — which is itself the well-documented over-policing of minority neighbourhoods. Feed those patterns back into a system that directs future policing, and you get a self-fulfilling loop. More policing in already over-policed areas produces more arrests. The prediction "verifies." The next allocation of resources follows. Richardson et al. (2019) treat this as a structural problem, not a statistical one. More data of the same kind doesn't solve it.

6.2 Legal and Constitutional Dimensions

AI in criminal justice also raises questions that go well beyond accuracy. Due-process rights in any functioning democratic legal system include the right to know the basis of a decision that restricts your liberty, and the right to challenge that reasoning meaningfully. Both rights become very difficult to exercise when the basis is a proprietary algorithm whose parameters are commercial secrets, or whose mathematical workings exceed what anyone other than a specialist can interpret. *Loomis v. Wisconsin* (2016) put this question directly before a court when the defendant challenged the use of a COMPAS score in sentencing on due-process grounds. The Wisconsin Supreme Court upheld the practice while

acknowledging the transparency concern — a result that has drawn sustained criticism from legal scholars since.

The point this raises is structural. The whole architecture of criminal procedure — adversarial proceedings, the right to confront evidence, the standard of proof, appellate review — assumes that decisions are made by reasoning agents whose logic can be examined, contested, and revised. Current AI systems cannot fill that function, no matter how accurate their predictions get.

7. AI Decision-Making in Recruitment and Human Resources

7.1 Volume Argument

Recruitment has adopted AI quickly, and for a straightforward reason. Large employers handle thousands of applications per role, and traditional human review doesn't scale cost-effectively to that volume. Applicant Tracking Systems that filter on keywords, GPA cut-offs, and similar criteria have been around for years. Newer tools add machine-learning layers on top — analysing video interviews for facial expression and vocal tone, scoring writing samples for inferred personality, ranking candidates on predicted performance from behavioural data.

Unilever and Hilton have publicly described their use of AI video analysis in initial screening, reporting meaningful drops in time-to-hire. The efficiency case is genuine. Consistent scoring criteria. Lighter cognitive load on hiring managers. Faster throughput at the top of the funnel. For large Indian employers in IT services or BFSI, where graduate recruitment runs into tens of thousands of applications per cycle, the appeal is obvious.

7.2 Bias, Validity, and Legal Concerns

The performance evidence is messier than the vendor pitches suggest. Amazon's 2018 decision to scrap its AI recruitment tool — after discovering it was systematically downgrading resumes tied to women's organisations and colleges — is the famous example, but it is not isolated. Studies of resume-screening algorithms have repeatedly found gender, racial, and class biases inherited from historical hiring data. Name-based bias, where applicants with names associated with specific ethnic backgrounds receive systematically lower evaluations, has been documented in both human and algorithmic screening since Bertrand and Mullainathan (2004). AI systems trained on those decisions tend to copy the pattern, not correct it.

Validity is a separate concern, especially for the newer assessment formats. The American Psychological Association and other professional bodies have raised serious doubts about whether tools that analyse facial expression or vocal pattern actually measure anything that predicts job performance — or whether they are picking up cultural familiarity, accent, and expressiveness, which correlate with class background rather than competence. Illinois responded with the Artificial Intelligence Video Interview Act (2020), which requires applicant consent and bias auditing from employers who use such tools.

7.3 The Case for Human Oversight

Hiring decisions have features that make the case for meaningful human oversight strong. They are consequential for the individual — one outcome can shift someone's earning trajectory by years. They depend on contextual interpretation that AI currently handles poorly: why a non-linear career path might reflect resilience rather than instability, or why a candidate's communication style may reflect culture rather than competence. And they involve normative trade-offs — between predicted performance and diversity, between formal credentials and visible potential — that are ultimately value judgments rather

than empirical findings.

8. AI Decision-Making in Education

8.1 Applications in Educational Assessment and Support

Education is often framed as a place where AI saves time and personalises learning at a scale no single teacher can match. Adaptive learning platforms like Knewton and Carnegie Learning adjust difficulty and pacing in real time. Automated Essay Scoring systems grade written work using natural language processing. Early-warning systems try to flag students who may disengage before the disengagement becomes withdrawal. Indian ed-tech platforms have built versions of all three.

The practical value is real. A teacher with thirty students cannot continuously track each one's engagement and comprehension. A dashboard that surfaces patterns — declining attendance, drop in assignment quality, missed deadlines stacking up — gives a teacher information that would otherwise be hard to assemble in time to act on. Research on adaptive learning suggests it does help, particularly for students who benefit from immediate feedback and extra practice.

8.2 The Limits of Algorithmic Education

Education's aims, though, aren't reducible to measurable performance outcomes, and that is where the algorithmic approach hits its real limits. Teaching is relational in a way that diagnosis or fraud detection isn't. A teacher who notices that a strong student has gone quiet over three weeks, and who guesses correctly that something at home has shifted, is doing something that current AI doesn't replicate. A mentor who hands a student a difficult text not because a metric flagged readiness but because the student needs to be trusted and stretched is making a pedagogical bet that no optimisation function will reproduce.

Automated grading raises equity concerns of its own. Studies have shown that automated essay scoring rewards particular stylistic features — longer responses, Latinate vocabulary, syntactic complexity — that correlate with socio-economic advantage. A student who writes clearly but plainly may receive a lower score, not because the writing is weaker by any considered standard, but because it does not match the patterns the training data learned to reward (Perelman, 2012). In a country where access to formal English instruction varies sharply by region and class, that mismatch is not trivial.

8.3 Governance in Educational AI

Educational decisions also extend well beyond any single student. Curriculum design, assessment standards, resource allocation, and institutional culture shape what a society's next generation knows and values. Those are political and ethical decisions in the fullest sense. They require deliberation and human accountability. Handing them over to algorithmic systems without serious oversight would amount to abdicating the responsibility democratic societies accept for how they educate the people who come next.

9. Cross-Domain Findings and a Decision Taxonomy

9.1 Common Patterns Across Domains

Read across all five domains, a consistent picture emerges. AI performs reliably when tasks are well-defined, data-rich, and centred on pattern recognition. It struggles, more consistently than is often admitted, with moral reasoning, with cases that fall outside its training distribution, and with the kind of contextual nuance that experienced practitioners handle intuitively. Bias from historically unequal data

shows up across every sector studied, not just one or two.

The accountability gap is also consistent. Assigning responsibility when an AI system causes harm is genuinely difficult, regardless of the domain. Was it the engineers who built the system? The organisation that deployed it? The regulator that approved it? The answer is rarely clean, and legal frameworks for resolving it remain underdeveloped almost everywhere.

The strongest arguments for human oversight, across all five domains, cluster around the same three factors. High consequence — errors carry severe costs for the people affected. Moral complexity — the decision involves value trade-offs that resist algorithmic resolution. And limited reversibility — the decision is hard to undo once made. Where the stakes are low, the optimisation target is clean, and mistakes can be corrected easily, AI autonomy makes more sense.

9.2 A Tiered Decision Taxonomy

Drawing on the cross-domain analysis, this paper proposes a three-tier taxonomy for classifying decisions according to how much human oversight is warranted. The taxonomy isn't intended as a rigid rule. It is a way of organising a practical question policymakers and practitioners already face: where does AI belong, and where shouldn't it go alone?

Tier	Consequence Severity	Moral/Contextual Complexity	Reversibility	Recommended Model
Tier 1: AI-Assisted (Human Decides)	High — life, liberty, livelihood at stake	High — moral agency indispensable	Low — irreversible or near-so	Human makes final decision; AI provides analysis and flags
Tier 2: Collaborative (Human Reviews)	Medium — significant but bounded impact	Medium — context matters; criteria clearer	Medium — reversible with effort	AI recommends; human reviews and confirms before action
Tier 3: AI-Autonomous (Human Audits)	Low — routine operational choices	Low — well-defined criteria, data-rich	High — easily corrected	AI decides; human audits periodically for drift and bias

Table 1: Proposed Three-Tier Decision Taxonomy for Human-AI Collaboration

Tier 1 decisions — criminal sentencing, complex clinical diagnoses, employment termination, child welfare — should always be made by an accountable human being. AI can inform them by pulling relevant data, flagging inconsistencies, or widening the option set. It should not decide them. Tier 2 decisions — initial credit assessments with a right of appeal, scheduling decisions

with significant downstream effects, early-stage candidate screening with mandatory human review — can be AI-recommended, but they require a human to confirm before anything actually happens. Tier 3 decisions — flagging a transaction for temporary block pending review, content recommendation, inventory routing — can run autonomously, provided someone audits the system regularly for drift and bias.

The taxonomy is not static. Where a particular decision belongs should be revisited as AI capabilities improve and as real-world performance evidence accumulates. A task that needs Tier 1 oversight today may shift toward Tier 2 if transparency and accuracy improve materially. The burden of proof for *downgrading* the level of oversight should be high, however — and especially so for decisions that fall disproportionately on people who already face structural disadvantage.

FIGURE 1: Human-AI Decision Spectrum

TIER 1	**TIER 2**	**TIER 3**
Human Decides	**Human Reviews**	**AI Autonomous**
AI Supports	*AI Recommends*	*Human Audits*
Criminal justice, clinical decisions, child welfare, major employment outcomes	Credit assessment, medical screening, job shortlisting, student risk flags	Fraud blocking, content recommendation, inventory management, routing

10. Ethical Dimensions of AI Decision-Making

10.1 Accountability and the Responsibility Gap

One of the deeper ethical problems AI introduces is what Matthias (2004) called the "responsibility gap" — the difficulty of assigning moral and legal responsibility when an automated system causes harm. When Amazon's recruiting algorithm produced discriminatory outcomes, who was responsible? The engineers? The leadership that approved deployment? The organisation as a whole? The question matters in practice because it determines who can be held to account, what remedies are available, and what institutional incentives exist to prevent the same failure from happening again.

The gap is partly technical — system opacity makes it hard to trace how a specific outcome emerged from a model — and partly a governance failure. Most legal systems were not designed for cases where the proximate cause of a harmful decision is software rather than a person. The EU AI Act is the most ambitious regulatory attempt to date to close that gap, with explicit liability provisions and mandatory human oversight requirements for high-risk applications. India's Digital Personal Data Protection Act (2023) touches some of this ground for automated decisions involving personal data, though its specific obligations around AI are still being worked out through subsidiary rules.

10.2 Autonomy and Dignity

A related concern is autonomy. When a person receives a negative credit decision from a human loan officer, they can ask why, push back on the reasoning, and appeal. When the same decision comes from an algorithm — particularly one that's proprietary or mathematically complex — meaningful contestation becomes much harder. That difficulty is a harm in itself, separate from any question about whether the AI got the answer right. It treats people as objects of computation rather than agents who are owed reasons for the choices that shape their lives.

The GDPR's "right to explanation" in automated decision-making (Article 22) was an early attempt to address this, though its practical implementation remains contested. The underlying principle

generalises. People are owed a comprehensible account of consequential automated decisions. Explanation capacity should be designed into AI systems from the start, not bolted on later if there's time.

10.3 Equity and Distributive Justice

Across all five domains examined here, AI bias falls disproportionately on populations already marginalised. This isn't coincidence. Systems trained on historical data inherit the inequities embedded in that data, and those inequities have always fallen hardest on people of colour, women, lower-income groups, and others who have faced systematic discrimination. Deploying AI at scale on top of those patterns doesn't just reproduce historical injustice. It institutionalises it through technically sophisticated and often opaque means.

Addressing this requires more than better debiasing techniques, though those certainly matter. It needs human oversight that draws on substantive knowledge of social history and structural inequality — knowledge that can't be extracted from data alone. Audit bodies need diverse membership and real authority to require remediation when bias surfaces. Organisations deploying AI in high-stakes contexts should bear the burden of demonstrating fairness, rather than placing the burden of proving discrimination on the people most likely to be harmed by it.

11. Conclusion and Recommendations

11.1 Summary of Findings

This paper has worked through AI decision-making across five domains — healthcare, finance, law enforcement, recruitment, and education — and found consistent patterns in both capability and limitation. AI genuinely adds value in pattern recognition, data processing, and operational efficiency. It also struggles, more often and more consistently than the marketing literature admits, with contextual understanding, moral reasoning, and the structural requirements of accountability.

The three-tier taxonomy proposed in Section 9 is the practical response. Tier 1 decisions — high-stakes, morally complex, and largely irreversible — require accountable human judgment, with AI in a support role. Tier 2 decisions — significant but more bounded — should be AI-recommended and human-reviewed before they take effect. Tier 3 decisions — routine, low-stakes, and reversible — can be AI-autonomous, subject to ongoing audit.

11.2 Recommendations

A few recommendations follow. Policymakers and regulators should adopt risk-stratified oversight frameworks — versions of the three-tier taxonomy proposed here — as the basis for AI governance across sectors. High-risk applications should require mandatory pre-deployment bias audits, ongoing monitoring, and independent third-party oversight. Organisations deploying AI in consequential decision contexts should invest in explainability infrastructure that makes contestation actually possible for the people affected.

AI developers and researchers should treat transparency, interpretability, and bias mitigation as core engineering values, not compliance boxes to be ticked at minimum cost. The default position should be that a new AI decision system is biased until evidence shows otherwise, not the reverse. Higher education institutions — including the ones training the engineers who will go on to build these systems — should make sure their graduates leave with a real understanding of AI capabilities and limits, so that they can supervise the technology meaningfully rather than rubber-stamp it.

The wider cultural shift this paper argues for is to stop treating human oversight as friction to be

engineered away. The efficiency gains from AI are real and worth taking seriously. But efficiency isn't the only thing that matters when decisions affect people's lives. Accountability, fairness, and dignity aren't inefficiencies. They are conditions of legitimate governance, and building AI systems that actually serve those values is the harder, more consequential work ahead.

11.3 Limitations and Future Research

A few limitations are worth naming. The analysis here rests on secondary data and documented cases. The conclusions are grounded in a large literature, but they don't carry the precision of primary quantitative research. The five domains examined are representative, not exhaustive — governance, infrastructure management, environmental regulation, and social services would each repay similar treatment. The taxonomy itself needs empirical validation in real organisational settings, and the pace of AI development means specific performance findings may be overtaken by technical advances faster than the structural arguments will be.

Future research should operationalise the tier taxonomy for specific contexts, test whether applying it produces defensible outcomes in practice, and look at the political economy of AI governance — understanding why organisations adopt or resist oversight is at least as important as designing the oversight itself. Cross-national comparative work would also be valuable, given how much regulatory approaches vary across jurisdictions. Indian regulators are at a relatively early stage in this work, and there is an opportunity for domestic research to shape the rules before they harden.

References

1. Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There is software used across the country to predict future criminals. And it is biased against Blacks. ProPublica. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
2. Bertrand, M., and Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American Economic Review, 94(4), 991–1013.
3. Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data and Society, 3(1), 1–12.
4. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.
5. Corkery, M., and Protess, B. (2017). How JPMorgan Chase has used technology to cut lawyers' work. New York Times. Available at: https://www.nytimes.com
6. De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., and Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24(9), 1342–1350.
7. Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. ArXiv Preprint. arXiv:1702.08608.
8. Dressel, J., and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580.
9. European Commission. (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Brussels: European Commission.
10. Kahneman, D. (2011). Thinking, Fast and Slow. New York: Farrar, Straus and Giroux.

11. Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6(3), 175–183.
12. NITI Aayog. (2018). National strategy for artificial intelligence. New Delhi: Government of India.
13. Obermeyer, Z., and Emanuel, E. J. (2016). Predicting the future — big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219.
14. Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.
15. OECD. (2019). OECD Principles on AI. Paris: OECD Publishing. Available at: https:// www.oecd.org/going-digital/ai/principles/
16. Perelman, L. (2012). Critique of mark my words: National assessment of educational progress versus automated essay scoring. International Examination, 18(1), 1–17.
17. Richardson, R., Schultz, J. M., and Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review, 94, 192–233.
18. Ross, C., and Swetlitz, I. (2018). IBM pitched its Watson supercomputer as a revolution in cancer care. It's flopping. STAT News.
19. Stevenson, M. T., and Doleac, J. L. (2019). Algorithmic risk assessment in the hands of humans. Social Science Research Network. Available at SSRN: https://ssrn.com/ abstract=3489440
20. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.
21. Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124–1131.
22. Tyler, T. R. (2003). Procedural justice, legitimacy, and the effective rule of law. Crime and Justice, 30, 283–357.
23. Wilson, H. J., and Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. Harvard Business Review, 96(4), 114–123.
24. *Abhay Chauhan | 2210991143 | Chitkara University*