Lyceum India Journal of Social Sciences || ISSN (Online): 3048-6513
Volume: 3 Issue: 6 || April 2026 (Special Issue 2) || DOI: 10.5281/zenodo.19955543

Ethics of Artificial Intelligence

Mr. Sanket Revankar[1] & Prof. Sanjyot Pai Vernekar[2]

Abstract:

Artificial Intelligence (Al) has rapidly become integral to decision-making systems across domains such as hiring, finance, education, and law enforcement. While Al promises efficiency and scalability, it also introduces several ethical challenges. This paper examines the problem of algorithmic bias, focusing on how Al systems could amplify structural inequalities embedded in historical data. Using the Amazon hiring algorithm (2018) as a central case study, the paper demonstrates that Al tools are not impartial but socially situated technologies. It explores demographic parity, equal opportunity, and merit-based fairness in AI resource allocation. The paper also addresses issues of opacity, accountability, and loss of human agency associated with Al systems. It concludes that without fairness-aware design, transparency, and regulatory intervention, Al risks automating discrimination at scale while undermining democratic principles and natural justice.

Keywords: Artificial Intelligence, AI, Ethics, Society

Introduction

Artificial Intelligence (Al) today encompasses a variety of computational systems capable of performing tasks. It is gaining widespread use in socially significant domains such as hiring, credit scoring, facial recognition, predictive policing, and scholarship allocation. While Al systems promise efficiency, consistency, and scalability, they also pose significant ethical concerns. An assumption surrounding Al is its objectivity and neutrality. However, this assumption is deeply misleading. Al systems learn from historical data, and such data often reflects social inequalities, including biases based on gender, caste, race, and class. As a result, Al systems may reproduce and even amplify these inequalities, leading to discriminatory outcomes. This paper explores the ethical challenges posed by Al, focusing on algorithmic bias, fairness, and accountability. It raises a central question: Can Al systems be both efficient and just, or do they privilege optimization over equity?

[1] Research Scholar & [2] Dean SSPIS, Goa University

Algorithmic Bias: Nature and Origins

Algorithmic bias refers to unfair discrimination embedded within Al systems. Contrary to popular belief, bias in Al is not accidental but often emerges from the very structure of data and model design. The training data being historical data reflects past inequalities and model objectives like optimization goals (e.g., profit maximization) which may privilege efficiency over fairness.

Case Study: Amazon Hiring Algorithm

A landmark example of algorithmic bias is Amazon's experimental hiring tool developed around 2018. The system was designed to evaluate resumes and assign scores to job applicants. However, it was later abandoned after it demonstrated systematic gender bias. The model was trained on resumes submitted over a ten-year period; a dataset heavily skewed toward male applicants due to the gender imbalance in the technology sector. Consequently, the Al system learned to associate male characteristics with higher competence. One solution is to use diverse and representative datasets along with regular audit of models for bias. Another solution could be adjusting outcomes based on demographic representation. However, such interventions raise a fundamental ethical dilemma: Should fairness override merit?

Competing Conceptions of Fairness

One of the key challenges in Al ethics is the lack of a universally accepted definition of fairness. Fairness is not a singular or stable concept; rather, it is plural, context-dependent, and often contested. In algorithmic systems, this leads to conflicting outcomes, making it impossible to satisfy all fairness criteria simultaneously. This creates both a technical and ethical dilemma, as designers must choose which notion of fairness to prioritize Demographic parity, or statistical parity, is one amongst the most widely discussed fairness metrics in Al. It requires that outcomes be distributed equally across different social groups, regardless of their underlying characteristics. In the context of hiring, for example, demographic parity would mean that the proportion of selected candidates is the same across categories such as gender, caste, or race.

Advantages

Demographic parity plays a crucial role in preventing systematic exclusion. By ensuring equal representation in outcomes, it directly addresses historical patterns of marginalization and discrimination.

It is relatively simple to employ and transparent, making it attractive from both a technical and policy standpoint.

Limitations

However, demographic parity faces significant criticism. Its primary limitation lies in its indifference to differences in qualifications or performance among candidates. By focusing solely on equal outcomes, it may require selecting individuals in ways that appear to conflict with merit-based criteria. This highlight concerns about fairness at the individual level, where candidates may be treated unequally despite differing capabilities. Moreover, critics argue that demographic parity can lead to "quota-like" systems, which may be perceived as unjust or politically contentious. It also risks oversimplifying complex social realities by reducing fairness to numerical equality, without addressing deeper structural inequalities.

Structural Inequality and Al

Al systems do not operate in a social vacuum; rather, they are situated within and shaped by existing social, economic, and historical structures. Al does not generate inequality ex nihilo but reflects, and often amplifies pre-existing structural disparities. The outcomes produced by Al systems are therefore inseparable from the conditions under which their training data is generated. Historical patterns of discrimination; whether based on gender, caste, race, or class become encoded in datasets; which are then used to train Al systems, which learn and replicate these patterns under the guise of predictive accuracy. Questions such as "Who deserves an opportunity?", which inherently require ethical reasoning, are reframed as "Who has the highest score?". This is a seemingly objective and quantifiable metric. In doing so, Al obscures the value-based nature of decision-making processes and displaces human judgment with algorithmic outputs.

Such a transformation risks depoliticizing and de-ethicalising decisions that are inherently social and moral. It creates an illusion of neutrality. As a result, Al-driven decision-making may inadvertently legitimize existing inequalities by presenting them as outcomes of impartial computation rather than products of historical structures.

Proxy Discrimination and Allocative Harm

Algorithmic decision-making presents notable ethical challenges not only through bias but also through more subtle mechanisms such as proxy discrimination and allocative harm. These forms of injustice are particularly difficult to identify and challenge, as they often operate beneath the surface of neutral systems.

Proxy Discrimination

A common strategy to mitigate bias in Al systems is the removal of sensitive attributes such as gender, caste, or race from datasets. However, this approach is often insufficient. Al systems can still infer these attributes indirectly through correlated variables; commonly referred to as proxies-such as location, educational background, language, or income level. For instance, in the Indian context, variables such as residential area or type of schooling may strongly correlate with caste. Even if caste is explicitly excluded, the system may reconstruct it through these proxies, thereby reproducing discriminatory patterns. This phenomenon is known as proxy discrimination

The ethical concern here is twofold. First, discrimination becomes less visible and harder to detect, as it is no longer tied to explicit categories. Second, it undermines the assumption that "fairness through unawareness"; the idea that ignoring sensitive attributes ensures fairness is sufficient. In reality, such approaches may merely obscure bias rather than eliminate it, making algorithmic discrimination opaquer and more difficult to contest.

Allocative Harm

Allocative harm refers to the unjust distribution of resources, opportunities, or benefits by algorithmic systems. When Al is used to optimize efficiency, such as minimizing financial risk or maximizing returns; it may systematically disadvantage already marginalized groups.

Consider the case of a free skill training scheme allocation. An Al system trained on historical data may identify that individuals from certain marginalized communities have higher dropout rates. Interpreting this as a financial risk, the system may reduce the likelihood of allocating opportunities to such individuals. While this may seem reasonable from an efficiency standpoint, it ignores the structural causes underlying these patterns, such as poverty, discrimination, lack of institutional support, or hostile training environments.

Thus, the system effectively penalizes individuals for conditions beyond their control, reinforcing cycles of disadvantage. Automated decision-making systems often disproportionately burden vulnerable populations, transforming social inequalities into data-driven justifications for exclusion.

The Indian Context: Caste and Substantive Equality

Any discussion of fairness in Al must be situated within its specific socio-legal context. In India, the principle of equality is not limited to formal equality; treating everyone the same but extends to substantive equality, which seeks to address historical and structural disadvantages. This commitment is embedded in constitutional provisions such as Articles 14, 15, and 16, which not only guarantee equality before the law but also permit affirmative action for socially and educationally disadvantaged groups.

Caste, as a deeply entrenched axis of social stratification, remains a constitutionally recognized category for corrective justice. Policies such as reservations in education and employment are designed to counteract centuries of systemic exclusion. In this context, the idea that fairness can be achieved by simply ignoring sensitive attributes; often described as "fairness through unawareness", is fundamentally inadequate. If Al systems are designed to be "blind" to caste, they risk reinforcing existing inequalities rather than mitigating them. This is because caste-based disadvantage is not only an attribute but a structural condition that shapes access to resources, opportunities, and social capital. Ignoring caste in algorithmic decision-making may therefore reproduce patterns of exclusion under the guise of neutrality.

This challenges dominant approaches in Al ethics that equate fairness with the removal of sensitive variables. In the Indian context, such an approach may conflict with constitutional morality, which emphasizes corrective justice and substantive equality over formal neutrality. A more appropriate framework is that of fairness-aware Al, wherein personal data or social categories are not used for exclusion but for monitoring, auditing, and correcting bias. Such systems can identify whether certain groups are systematically disadvantaged and enable interventions to ensure equitable outcomes.

Loss of Agency and Natural Justice

The increasing reliance on Al systems in decision making processes has significant implications for individual agency and procedural fairness. As Al systems are deployed in domains such as employment, credit allocation, education, and welfare distribution; they increasingly determine access to essential resources that shape life opportunities. This shift raises critical ethical and legal concerns. Individuals subject to algorithmic decisions often experience a loss of agency, as they are unable to understand, influence, or meaningfully respond to outcomes that affect them. Unlike traditional decision-making processes, where human judgment allows for dialogue, and contextual consideration, Al systems typically operate through opaque computational mechanisms that provide little to no explanation for their outputs.

Three interrelated problems emerge in this context:

i. Erosion of Individual Agency: Individuals are reduced to data points or risk profiles, rather than being treated as autonomous moral agents capable of reasoning and self-determination.
ii. Inability to Challenge Decisions: Algorithmic opacity prevents individuals from identifying the basis of adverse decisions, thereby limiting their capacity to contest them.
iii. Absence of Appeal Mechanisms: Many Al-driven systems lack formal processes for review or human intervention, leaving affected individuals without recourse.

These developments pose a direct challenge to the principles of natural justice, which form the foundation of fair decision-making in legal and administrative systems. Two core principles are particularly relevant:

i. The Right to Explanation: Individuals have a moral and legal claim to know the rationale underlying decisions that impact them.
ii. The Right to Contest: Individuals must have the opportunity to challenge decisions and present their case.

Al systems that function as "black boxes" undermine both these principles. When decisions are presented as outputs of complex algorithms without transparent reasoning, they effectively place individuals in a position where they are subject to authority without accountability. From a philosophical perspective, this transformation signals a shift from deliberative justice to automated governance, where decisions are framed by technical systems and rendered inaccessible to those affected by them. The result is a form of epistemic asymmetry, in which institutions using AI wield decision-making power while individuals lack the knowledge and tools necessary to question or resist it.

Conclusion

This paper has demonstrated that Al systems are not neutral or purely technical tools, but socio-technical systems deeply embedded within existing structures of inequality. The analysis of the Amazon hiring algorithm demonstrates how Al can replicate and amplify historical biases, even in technologically advanced and well-resourced environments. Rather than eliminating human prejudice, Al often encodes and scales it, thereby transforming localized discrimination into systemic and automated injustice.

A series of critical insights emerge from this study. First, algorithmic bias is not incidental but structural, arising from historically situated data, model design, and optimization goals. Second, fairness in Al is not

a singular or universally agreed-upon concept; instead, it consists of multiple, often conflicting frameworks, such as demographic parity, equal opportunity, and merit-based fairness. These tensions reveal that ethical decision-making cannot be fully reduced to technical optimization.

Third, the assumption that fairness can be achieved by ignoring sensitive attributes is fundamentally flawed, particularly in contexts like India, where substantive equality requires active recognition of social difference. In such cases, fairness-aware Al must respect, rather than being blind to structural inequalities.

Finally, the paper has highlighted the pressing need for transparency, accountability, and procedural justice in Al systems. The opacity of algorithmic decision-making not only obscures bias but also undermines individual agency, limiting the ability of affected persons to understand, challenge, or appeal decisions that shape their lives.

Given these concerns, the ethical deployment of Al requires deliberate and sustained intervention. This consists of the use of heterogeneous and representative datasets, the incorporation of fairness constraints, continuous auditing, and robust regulatory frameworks. Equally important is the preservation of human oversight and the protection of fundamental rights such as explanation and contestability.

References

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732. https://doi.org/10.15779/Z38BG31

Belenguer, L. (2022). Al bias: Exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adaptable to law. Al and Ethics, 2, 711-728. https://doi.org/10.1007/s43681-021-00138-8

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of Machine Learning Research, 81, 1-11.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186. https://doi.org/10.1126/science.aal4230

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. Proceedings of the

Conference on Fairness, Accountability, and Transparency, 329-338. https://doi.org/10.1145/3287560.3287589

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. Proceedings of Innovations in Theoretical Computer Science. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in HR recruitment. Business Research, 13(3), 795-848. https://doi.org/10.1007/s40685-020-00134

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application, 8, 141-163. https://doi.org/10.1146/annurev-statistics-042720 -125902

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59-68. https://doi.org/10.1145/3287560.3287598

Simon, J., Wong, P.-H., & Rieder, G. (2020). Algorithmic bias and the value sensitive design approach. Internet Policy Review, 9(4). https://doi.org/10.14763/2020.4.1534

Tolan, S., Miron, M., Gómez, E., & Castillo, C. (2019). Why machine learning may lead to unfairness: Evidence and remedies. Big Data & Society, 6(2). https://doi.org/10.1177/2053951719843994

van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. Journal of Business Research, 144, 93-106. https://doi.org/10.1016/j.jbusres.2022.01.076

Varsha, P. S., & Sharma, R. (2023). How can we manage biases in artificial intelligence systems? Journal of Innovation & Knowledge, 8(2). https://doi.org/10.1016/j.jik.2023.100337