Artificial Intelligence: A Construct of Human Values (Exploring How Human Biases Are Reflected in AI)

IYANUOLUWA ALARAPE

Abstract- ***Artificial Intelligence (AI) is increasingly seen as a solution to societal challenges, praised for its efficiency, objectivity, and ability to make decisions free from human error. However, this perception obscures the reality that AI is deeply embedded with human values, reflecting the biases of its creators. This paper argues that AI systems, far from being neutral, are sociotechnical constructs influenced by the social, political, and cultural contexts in which they are developed. The article explores how human biases infiltrate AI through data selection, feature prioritization, and algorithmic design, often perpetuating systemic inequalities. It highlights real-world cases of biased AI outcomes across sectors such as healthcare, hiring, criminal justice, and education. Furthermore, the paper addresses the ethical and philosophical implications of these biases, examining the power dynamics that influence AI development and the need for more inclusive and accountable design practices. By proposing strategies for mitigating bias, such as diverse development teams, transparency, and ethical frameworks, the paper aims to foster the creation of AI systems that serve the public good while ensuring fairness and equity.***

Indexed Terms- ***Artificial Intelligence, Human Bias, Sociotechnical Systems, Ethical AI, Algorithmic Bias, Fairness in AI***

I. INTRODUCTION

Artificial Intelligence (AI) is frequently celebrated as a dispassionate and infallible technology capable of making decisions more efficiently and accurately than humans. This techno-optimistic narrative frames AI as neutral, objective, and detached from human flaws (Crawford, 2021). Governments, corporations, and institutions increasingly rely on AI for tasks ranging from resume screening and medical diagnostics to loan approvals and criminal sentencing. In many of these contexts, AI is perceived as a way to reduce human error, eliminate discrimination, and optimize outcomes. However, this perception conceals a more complex and often unsettling reality: AI is not an autonomous or value-free entity.

Rather, AI systems are deeply human constructs. They are developed, trained, and implemented by individuals and institutions embedded within social, political, and historical contexts. From the datasets used to train machine learning models to the metrics used to assess performance, AI reflects the values, assumptions, and power dynamics of its creators. As such, it is not immune to the systemic biases that exist in society it can absorb, reinforce, and even amplify them.

This article explores the central thesis that AI is not merely a technological artifact, but a sociotechnical construct a product of both technical design and human decision-making. We argue that understanding AI through this lens is crucial to developing systems that are fair, inclusive, and accountable.

To unpack this, we explore several critical dimensions:

1.1 The Illusion of Objectivity in AI

Despite popular belief, AI systems are not inherently neutral. The illusion of objectivity often stems from their mathematical sophistication or black-box nature. However, every AI system is built on subjective decisions: What data should be included? What outcomes should be optimized? What trade-offs are acceptable? These questions are shaped by cultural norms, economic interests, and institutional priorities. As a result, AI does not operate in a vacuum it mirrors the values of the environments in which it is developed.

1.2 Sociotechnical Framing of AI

Viewing AI as a sociotechnical system highlights that technology cannot be separated from the society that builds and uses it. This framing encourages us to go beyond technical fixes and examine how power, inequality, and historical context influence technological development. It pushes back against deterministic narratives that present AI as an unstoppable force, reminding us that we have agency in how AI is designed and deployed.

1.3 Human Bias as a Design Ingredient

Bias in AI is often described as a problem of "dirty data" or poor training examples. But bias can also enter through the choices engineers make choices about which features to prioritize, which problems to solve, and which groups to serve. These decisions often reflect broader social biases, whether intentional or unconscious. For example, prioritizing profit over fairness can result in algorithms that systematically disadvantage marginalized communities. Bias, then, is not an anomaly it is a design ingredient that must be intentionally addressed.

1.4 Power and Representation in AI Development

A critical but often overlooked issue is who gets to build AI, and for whom. The technology sector is still dominated by a relatively homogeneous group predominantly male, Western, and economically privileged. This lack of diversity influences how problems are defined, what solutions are pursued, and whose perspectives are ignored. AI systems are thus often built with blind spots that reflect the absence of marginalized voices in the design process. Representation is not just a matter of fairness it directly affects the integrity and social impact of AI systems.

1.5 Objectives of This Article

This article sets out to challenge the assumption that AI is a neutral tool. Instead, we present a comprehensive examination of how human values are encoded into AI systems. The paper is structured around the following core objectives:

- To analyze the human and historical origins of AI bias
- To examine real-world manifestations of these biases across sectors
- To assess the societal consequences of biased AI systems
- To explore the ethical and philosophical questions raised by AI decision-making
- To recommend actionable strategies for building more equitable and accountable AI systems

1.6 Significance of the Inquiry

The stakes are high. As AI becomes increasingly integrated into critical decision-making domains such as healthcare, finance, education, and criminal justice it has the power to shape life outcomes for individuals and communities. If these systems remain opaque and unregulated, they risk entrenching existing social inequities under a veneer of algorithmic legitimacy. Understanding AI as a reflection of human values, and not a substitute for them, is essential for ensuring that technology serves the public good.

II. THE HUMAN ORIGIN OF AI SYSTEMS

AI systems do not emerge in isolation they are designed, trained, and implemented within complex social, cultural, and political environments. Every aspect of AI development is infused with human judgment and decision-making. From the selection of datasets to the formulation of algorithms and the setting of optimization goals, each step reflects the priorities, perspectives, and limitations of human developers (Barocas & Selbst, 2016; Noble, 2018).

AI technologies are constructed from:

- Data sourced and labeled by humans – which may carry historical biases and social inequities.
- Algorithms programmed with subjective design choices – such as which features to include or ignore.
- Objectives selected based on human values – such as prioritizing profit, efficiency, or fairness.

Consequently, AI systems often replicate and even intensify systemic social inequalities. For instance, a recruitment tool trained on historical hiring data may unintentionally discriminate against women if the original dataset reflects a male-dominated industry (Dastin, 2018). Similarly, predictive policing algorithms based on legacy crime data can reinforce racial profiling by repeatedly flagging minority neighborhoods as high-risk (Richardson, Schultz, & Crawford, 2019). These outcomes highlight that AI systems are not impartial arbiters of truth they are deeply shaped by the social conditions and intentions of their creators.

2.1 Sources of Human Influence

Several interrelated mechanisms allow human bias to infiltrate AI systems:

2.1.1 Historical Bias in Data

Historical datasets often reflect longstanding inequities and discriminatory practices. When such data is used to train machine learning models, the AI "learns" those biases as if they are neutral truths. For example, facial recognition tools trained on predominantly lighter-skinned individuals perform poorly on darker-skinned faces, leading to higher error rates for minority groups (Buolamwini & Gebru, 2018). This is not a flaw in the algorithm's logic, but in the data it was fed.

2.1.2 Subjective Feature Selection

Developers make numerous choices about which features to include in a model. These decisions are rarely neutral. Choosing to emphasize certain variables such as zip code in credit scoring can serve as proxies for race or socioeconomic status, even if unintentionally (Barocas et al., 2019). Feature selection thus becomes a reflection of the developer's assumptions about what matters.

2.1.3 Value-Laden Optimization Metrics

AI systems are designed to optimize for specific goals such as accuracy, speed, or cost-effectiveness. However, these metrics are themselves value-laden. A healthcare algorithm that optimizes for cost savings might systematically deprioritize high-need patients from marginalized communities, not because of clinical irrelevance, but due to economic profiling (Obermeyer et al., 2019). Choosing which outcomes to optimize is inherently a moral and political decision.

These mechanisms make clear that human influence in AI is not incidental it is foundational. Understanding these origins is the first step toward building systems that are not just technically effective but socially responsible.

III. MANIFESTATIONS OF BIAS IN AI

Bias in artificial intelligence does not exist in a vacuum it manifests at various levels of the AI pipeline, from the construction of training data to the outputs produced during real-world deployment. These biases can cause measurable harms, particularly for already marginalized or vulnerable populations. Understanding how bias emerges and operates within AI systems is crucial for developing equitable and responsible technologies.

3.1 Dataset Bias

Dataset bias occurs when the training data used to build an AI system is not representative of the broader population it is intended to serve. This often results in models that perform well for certain groups but poorly for others. For instance, Buolamwini and Gebru (2018) found that commercial facial recognition systems had error rates of up to 34.7% for darker-skinned women, compared to less than 1% for lighter-skinned men. This discrepancy was largely due to the training datasets containing predominantly light-skinned, male faces.

Another example is language models trained primarily on English text or Western cultural contexts, which may misinterpret or marginalize non-Western vernaculars, idioms, and expressions (Bender et al., 2021). Such biases can lead to inaccurate translations, culturally insensitive outputs, and other forms of exclusion.

Bias at the dataset level often stems from:

- Historical exclusion of marginalized groups from data collection processes.
- Overrepresentation of specific geographies, demographics, or viewpoints.
- Implicit labeling bias during data annotation.

3.2 Algorithmic Bias

Algorithmic bias arises from design decisions made during the development of machine learning models. Even when datasets are balanced, the algorithms themselves may encode bias based on the features selected, the weights assigned, or the optimization objectives defined.

For example, employment screening algorithms may de-prioritize applicants with non-linear career paths, such as those who took parental leave or changed careers reflecting a narrow, traditional view of career success (Raji et al., 2022). Similarly, healthcare algorithms that optimize for cost-effectiveness may inadvertently reduce access to care for high-risk but economically disadvantaged patients (Obermeyer et al., 2019).

These biases reflect the implicit values and assumptions of developers, often privileging efficiency, profit, or risk minimization over fairness, equity, and inclusivity.

3.3 Feedback Loop Bias

A feedback loop bias, also known as *automation bias* or *runaway reinforcement*, occurs when the outputs of an AI system influence the environment in ways that reinforce the system's original assumptions often exacerbating existing inequalities.

A classic case is predictive policing. Ensign et al. (2018) describe how predictive algorithms trained on historical crime data often send more law enforcement to certain neighborhoods usually low-income or predominantly minority areas. The increased police presence leads to more recorded incidents (regardless of actual crime rates), which then reinforces the algorithm's belief that the area is high-risk. Over time, this creates a self-perpetuating cycle of surveillance and over-policing.

Feedback loop bias can also appear in online recommendation systems, where popular content is promoted more heavily, making it even more popular, while suppressing niche or minority content, leading to lack of visibility and representation.

3.4. The AI Mirror – Reflections of Human Values and Societal Norms

The book The AI Mirror by Prof. Shannon Vallor provides a profound exploration of how artificial intelligence (AI) systems are not just tools but reflections of human values, biases, and societal norms. According to Shannon, AI acts as a "mirror," reflecting the priorities, ethics, and cultural contexts of its creators. This perspective challenges the notion that AI is neutral or objective, emphasizing instead that AI systems embody the intentions and limitations of the humans who design them. By examining this concept, we can better understand how AI perpetuates or disrupts existing social structures and how it can be designed to align with ethical principles.

Key Insights from The AI Mirror

AI as a Reflection of Human Intentions

Shannon argues that AI systems are deeply influenced by the data they are trained on and the goals set by their developers. For example, if an AI system is designed to optimize efficiency in hiring processes, it may inadvertently prioritize candidates from specific demographic groups based on historical hiring patterns, thus reinforcing systemic biases. This highlights the importance of intentionally embedding fairness and inclusivity into AI design.

The Role of Values in AI Development

The book underscores the need for value-driven AI development. Shannon advocates for frameworks such as the IEEE's Ethically Aligned Design, which emphasizes aligning AI systems with human well-being, transparency, and accountability. These frameworks ensure that AI serves as a tool for enhancing societal equity rather than entrenching inequalities.

AI and the Amplification of Societal Biases

A central theme in The AI Mirror is the risk of AI amplifying societal biases. For instance, facial recognition technologies have been shown to perform poorly for women and people of color due to non-diverse training datasets. Shannon calls for proactive measures to mitigate these risks, such as auditing datasets for bias and involving diverse teams in AI development.

Human-Centric AI: Shifting the Paradigm

Shannon proposes a shift toward human-centric AI, where technology is designed to prioritize human dignity and well-being. This involves creating AI systems that are transparent, explainable, and aligned with ethical principles. For example, AI-powered healthcare systems should prioritize patient autonomy and informed consent over purely algorithmic decision-making.

IV. REAL-WORLD CASE STUDIES OF AI BIAS

The theoretical risks of bias in artificial intelligence (AI) are not speculative they have already produced tangible and often harmful outcomes across several sectors. From hiring and healthcare to criminal justice and education, AI systems have repeatedly demonstrated how algorithmic decisions can perpetuate, or even worsen, existing social inequalities when built on biased foundations. Below are key real-world examples where bias in AI systems has led to measurable harm, along with explanations of their root causes.

Domain	AI System	Biased Outcome	Root Cause
Hiring	Amazon Resume Screening Tool	Penalized women applying for technical roles	Trained on resumes from a male-dominated workforce
Healthcare	U.S. Risk Assessment Algorithm	Assigned lower risk scores to Black patients	Optimized for healthcare cost, not clinical need
Criminal Justice	COMPAS Recidivism Tool	Gave higher risk scores to Black defendants	Trained on racially biased arrest records
Facial Recognition	IBM, Microsoft, and others	High error rates for darker-skinned and female faces	Non-diverse training datasets
Education	UK COVID-19 Grading Algorithm	Lowered grades for students from disadvantaged schools	Relied on past institutional performance

4.1 Hiring: Amazon Resume Screening Tool

In 2018, Amazon discontinued an internal AI tool that had been used to automate the screening of job applicants. The system, trained on a decade's worth of resumes submitted to the company, learned to favor male candidates for technical roles downgrading applications that included the word "women's," such as "women's chess club captain" (Dastin, 2018). This outcome stemmed from the AI reflecting patterns in past hiring practices, which were overwhelmingly male-dominated. Although the algorithm was never deployed in live hiring, the case highlights how AI can replicate historical inequalities when training data is biased.

4.2 Healthcare: U.S. Risk Assessment Algorithms

A 2019 study revealed that a widely used healthcare algorithm in the U.S., which affected decisions for millions of patients, systematically assigned lower risk

scores to Black patients compared to white patients with the same level of health needs (Obermeyer et al., 2019). The algorithm's objective was to identify patients who would benefit from extra care programs, but it used healthcare cost as a proxy for need overlooking the fact that Black patients often receive less expensive care due to structural inequalities. As a result, many high-risk Black patients were not flagged for support.

4.3 Criminal Justice: COMPAS Recidivism Tool

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm is used in various U.S. courts to assess the likelihood that a defendant will reoffend. A 2016 investigation by ProPublica found that the tool was twice as likely to incorrectly label Black defendants as high risk compared to white defendants (Angwin et al., 2016). This disparity was traced to systemic racial bias in historical policing and arrest records, which formed the basis of the algorithm's training data. The COMPAS case underscores how biased datasets can encode and automate racial injustice.

4.4 Facial Recognition: Commercial AI Tools

A study by Buolamwini and Gebru (2018) examined commercial facial analysis systems developed by companies like IBM, Microsoft, and Face++. It found that these systems had an error rate of 0.8% for lighter-skinned men but up to 34.7% for darker-skinned women. These disparities were linked to the underrepresentation of dark-skinned individuals in the training datasets. Facial recognition bias has significant real-world implications ranging from misidentification in law enforcement to exclusion in digital identity systems.

4.5 Education: UK COVID-19 Grading Algorithm

During the COVID-19 pandemic, the UK government used an AI grading algorithm to assign student scores when final exams were cancelled. The algorithm penalized students from schools with historically lower performance typically underfunded public schools in economically disadvantaged areas while favoring those from elite institutions. The model's reliance on school-level performance rather than individual student ability caused widespread outrage and was eventually scrapped after public backlash (Smith, 2020). This case illustrates how institutional bias can be codified and enforced by algorithmic systems.

V. SOCIETAL CONSEQUENCES OF AI BIAS

Bias in artificial intelligence (AI) systems is not merely a technical flaw it can translate into deep, structural harms that perpetuate social inequality, discrimination, and injustice. Because AI systems are increasingly embedded in decision-making processes across sectors, the amplification of historical bias through these systems has tangible and sometimes devastating societal consequences.

When AI is deployed at scale, its outcomes can reinforce preexisting power imbalances, often under the veneer of technological neutrality and objectivity. These consequences are disproportionately borne by marginalized groups, further entrenching systemic disadvantage (Noble, 2018; Benjamin, 2019).

5.1 Employment Discrimination

Biased AI systems used in recruitment and employee evaluation can inadvertently filter out qualified candidates from underrepresented groups. For example, hiring algorithms trained on historical data from predominantly white or male applicants may penalize resumes that signal gender, ethnicity, or educational background associated with marginalized populations (Raghavan et al., 2020).

Even seemingly neutral factors like gaps in employment history can result in biased outcomes disadvantaging candidates who took time off for caregiving or illness. The result is a feedback loop that perpetuates exclusion from economic opportunities under the guise of meritocracy.

5.2 Healthcare Inequity

AI-driven diagnostic tools and risk prediction algorithms are increasingly used to allocate medical resources. However, these tools can reflect and

reinforce disparities in access to healthcare. As shown by Obermeyer et al. (2019), some healthcare algorithms use healthcare spending as a proxy for need failing to account for structural inequities that limit care access in Black and low-income communities.

This can lead to underdiagnosis or neglect of vulnerable patients, reducing their chances of early intervention or life-saving treatment. AI systems that fail to account for these disparities may exacerbate, rather than alleviate, health inequalities.

5.3 Over-Policing and Criminal Justice Disparities

Predictive policing tools that rely on crime data are particularly susceptible to bias, as they often reflect discriminatory patterns in past law enforcement activity. Communities that have been historically over-policed particularly Black and Latino neighborhoods become labeled as high risk by these systems (Lum & Isaac, 2016).

This creates a self-fulfilling prophecy: more patrols lead to more arrests, which reinforce the dataset used to predict future crime. AI thereby acts as a mechanism of surveillance that disproportionately targets marginalized populations, while reinforcing the false legitimacy of biased data.

5.4 Educational Inequality

AI systems used for grading or admissions have also been shown to disadvantage students from underprivileged backgrounds. A notable example is the UK's 2020 COVID-19 grading algorithm, which assigned lower scores to students from schools in lower-income areas (Smith, 2020). The model relied on historical school performance rather than individual student potential.

Such practices undermine educational equity and restrict social mobility, penalizing students not for their abilities but for their socioeconomic status. In doing so, AI-driven educational tools risk further entrenching generational disadvantage.

5.5 Institutional Trust and Social Legitimacy

When AI systems consistently yield unfair outcomes, they can erode public trust in institutions be they employers, hospitals, courts, or schools. A lack of transparency and accountability in AI decision-making exacerbates this mistrust, especially when affected individuals have no clear means to challenge or appeal those decisions (Pasquale, 2015).

The danger is that biased AI systems not only produce discriminatory outcomes, but also legitimize them through a veneer of computational "objectivity." As Ruha Benjamin (2019) cautions, this can result in "the New Jim Code" the automation of inequality through design choices that appear race-neutral but are anything but.

VI. ETHICAL AND PHILOSOPHICAL CONSIDERATIONS

As artificial intelligence (AI) becomes more deeply embedded in decision-making processes, it raises profound ethical and philosophical questions that go far beyond concerns of technical performance. AI does not simply automate tasks it automates judgments. In doing so, it encodes values, priorities, and worldviews that may not be transparent or universally shared.

Critically, AI forces us to ask not just what we *can* do with intelligent systems, but what we *should* do.

6.1 Whose Values Are Embedded in AI?

AI systems are built on normative assumptions about what constitutes success, fairness, harm, or benefit. These assumptions are typically embedded by developers, who may not be representative of the diverse societies their technologies affect (Crawford, 2021). As a result, AI systems often reflect the perspectives of dominant cultural, economic, or political groups.

For example, facial recognition algorithms have been developed largely using datasets sourced from Western populations, which results in higher error rates when applied to non-Western faces (Buolamwini & Gebru, 2018). This is not just a technical flaw it is a

reflection of whose identities and experiences are considered in AI design, and whose are excluded.

6.2 Defining Fairness: Whose Standard?

The notion of "fairness" in AI is not a fixed concept. Multiple definitions exist such as statistical parity, equal opportunity, or individual fairness and these definitions can conflict (Binns, 2018). Choosing one definition over another is inherently a political and ethical decision, yet it is often treated as a technical problem.

This raises critical questions: Who decides what kind of fairness matters? What trade-offs are being made, and who bears the cost? Without public accountability, these choices may reinforce the interests of powerful institutions rather than vulnerable groups.

6.3 Can Machines Be Trusted to Make Moral Judgments?

Delegating value-laden decisions to AI such as who gets bail, which job applicants are short-listed, or how scarce resources are distributed risks distancing ethical accountability from human actors. Unlike human decision-makers, machines lack moral agency, empathy, or contextual understanding (Mittelstadt et al., 2016).

While AI can be programmed to simulate ethical reasoning, it cannot truly *understand* moral complexity, cultural nuance, or human emotion. Relying on AI to make such decisions raises fundamental concerns about accountability, responsibility, and justice.

6.4 Beyond Technical Fixes: Toward Ethical Governance

Efforts to reduce bias in AI systems often focus on technical debiasing correcting datasets, adjusting models, or improving fairness metrics. While important, such approaches are insufficient on their own. Ethical AI development must include a broader conversation about the social purposes of technology and the power dynamics it reinforces (Floridi et al., 2018).

This requires an interdisciplinary approach, involving:

- Technologists, who build and deploy the systems.
- Ethicists and philosophers, who reflect on the values and assumptions embedded in design.
- Legal scholars, who explore regulatory and rights-based implications.
- Community members, particularly those most affected, who must have a voice in how technologies are developed and used.

Participatory design and inclusive governance structures are essential to ensure AI reflects diverse moral perspectives and protects the rights and dignity of all people.

6.5 Ethical Frameworks in Practice

Several organizations and institutions have developed ethical frameworks to guide responsible AI design, such as:

- OECD Principles on AI (2019): Emphasizing transparency, accountability, and human-centered values.
- EU Ethics Guidelines for Trustworthy AI (2019): Including principles like beneficence, non-maleficence, and explicability.
- IEEE Ethically Aligned Design: Advocating for human well-being and transparency.

These frameworks provide useful starting points, but their implementation must be accompanied by institutional will, legal backing, and cultural change.

VII. MITIGATING BIAS IN AI SYSTEMS

As the ethical concerns surrounding artificial intelligence (AI) grow, researchers and practitioners are increasingly focusing on how to reduce bias and promote fairness in AI design and deployment. Effective mitigation requires a comprehensive, multi-stakeholder approach that integrates technical solutions, organizational practices, regulatory mechanisms, and participatory governance.

7.1 Inclusive and Diverse Development Teams

Diverse AI teams across gender, race, class, geography, and expertise are more likely to anticipate the real-world impacts of AI systems on marginalized communities (West et al., 2019). Inclusive teams bring varied lived experiences and cognitive perspectives, which helps challenge assumptions and reduce blind spots in system design.

Aspect	Biased Outcome	Debiased Outcome
Loan Approval	Favors affluent applicants	Applies fairness constraints for demographic equity
Recruitment	Penalizes employment gaps (e.g., parental leave)	Context-aware candidate evaluation
Predictive Policing	Over-polices minority neighborhoods	Balances data with socioeconomic and demographic context
Medical Diagnosis	Lower accuracy for minority patients	Training data includes diverse patient populations
Student Scoring	Penalizes schools in low-income areas	Adjusts grading based on resource disparities

By embedding diversity into every stage of the development pipeline from data collection to model evaluation organizations can better align AI systems with the values of the communities they serve.

7.2 Bias Audits and Algorithmic Transparency

Bias in AI is not always visible at first glance. Periodic algorithmic audits are essential to uncover and address systemic inequities. These audits should evaluate inputs, outputs, decision pathways, and feedback loops.

Explainable AI (XAI) further enhances accountability by enabling users to understand how and why a decision was made (Doshi-Velez & Kim, 2017). Transparency promotes trust, empowers users to challenge unjust outcomes, and facilitates third-party review.

7.3 Regulatory and Ethical Frameworks

In addition to technical improvements, robust regulatory oversight is vital. Policymakers, standards bodies, and ethics boards have a responsibility to ensure that AI systems operate in a fair and accountable manner. Key initiatives include:

- Datasheets for Datasets: Structured documentation that details the motivation, composition, and potential biases of datasets (Gebru et al., 2018).
- Algorithmic Impact Assessments: Risk evaluations that precede deployment of high-impact AI systems.
- Human-in-the-Loop Decision Systems: Ensuring that humans retain final decision-making power in contexts involving social or ethical judgment.

These measures provide formal mechanisms to hold developers and organizations accountable.

7.4 Community Participation

Participatory design practices aim to include those most affected by AI systems especially marginalized or historically excluded populations in the design and evaluation processes (Costanza-Chock, 2020). This democratizes AI development and helps ensure technologies promote equity rather than reinforce systemic harm.

Community workshops, citizen panels, and collaborative design sessions offer forums for lived experience to shape technical systems.

Sample Ethical Review Checklist

Category	Key Questions to Ask
Data Sourcing	Is the dataset representative? Was informed consent obtained?
Bias Detection	Have fairness audits been conducted? Are disparities across groups measured and addressed?
Transparency	Can decisions be explained to stakeholders? Are models open for external review?
Accountability	Who is responsible for errors or harm? Is redress available?
User Impact	Are vulnerable users protected? Does the system reinforce or reduce inequality?
Human Oversight	Is a human involved in key decisions? Can AI outcomes be contested or overridden?

CONCLUSION

Artificial Intelligence is not an autonomous force devoid of values. It is a reflection of human priorities, prejudices, and philosophies encoded in the data it learns from, the algorithms it runs on, and the institutional systems it operates within. When AI is treated as infallible, we risk reinforcing and amplifying systemic inequalities behind a façade of technical neutrality and sophistication.

This article has demonstrated that AI systems are sociotechnical constructs. Their development is shaped by historical data, subjective design choices, and embedded values. These human inputs can lead to real-world harm discriminatory hiring, unfair policing, inequitable healthcare, and biased education systems if left unchecked.

To build AI that is just, inclusive, and trustworthy, we must first recognize its human roots. This recognition should guide the development of policies, frameworks, and design processes that center equity, transparency, and accountability. The future of AI does not depend solely on better algorithms but on better questions, better ethics, and better collaboration.

8.1 Recommendations

Based on the findings and insights in this paper, the following recommendations are proposed for policymakers, developers, researchers, and civil society stakeholders:

1. Center Equity at Every Stage of AI Lifecycle

- Ensure fairness and inclusiveness from data collection to deployment.
- Regularly audit for disparate impacts on vulnerable or underrepresented groups.

2. Mandate Algorithmic Transparency

- Require explainable AI tools in critical applications such as healthcare, criminal justice, and education.
- Create public registers for high-impact AI systems with documented design choices, training data sources, and evaluation metrics.

3. Promote Interdisciplinary Collaboration

- Involve ethicists, sociologists, legal scholars, and affected communities in AI development processes.
- Fund research that explores the ethical, legal, and societal implications (ELSI) of AI technologies.

4. Build Diverse and Inclusive AI Teams

- Increase representation from historically marginalized groups in tech development and leadership roles.
- Implement anti-bias training and inclusive hiring practices within AI labs and companies.

5. Adopt and Enforce Regulatory Frameworks

- Governments should pass legislation for algorithmic accountability, such as:
    - Algorithmic Impact Assessments
    - Right to Explanation
    - Bias Audits and Risk Mitigation Plans

6. Support Public Participation and Literacy

- Create mechanisms for community input in the development and oversight of public AI systems.
- Educate the public about AI's limitations and potential harms, empowering them to challenge biased outcomes.

7. Encourage Open Research and Shared Datasets

- Incentivize the creation and sharing of high-quality, balanced, and well-documented datasets.
- Support open-source algorithmic tools that facilitate replicability, peer review, and transparency.

8. Embed Human Oversight and Contestability

- Retain meaningful human control in decision-making processes involving rights and well-being.
- Ensure accessible mechanisms for appeal, redress, and accountability in cases of harm.

REFERENCES

[1] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. http://fairmlbook.org

[3] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732. https://doi.org/10.2139/ssrn.2477899

[4] Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity Press.

[5] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

[6] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149–159. https://doi.org/10.1145/3287560.3287598

[7] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91.

[8] Costanza-Chock, S. (2020). *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press.

[9] Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

[10] Dastin, J. (2018, October 10). Amazon scrapped "sexist AI" recruiting tool. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[11] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. https://arxiv.org/abs/1702.08608

[12] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 160–171.

[13] Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

[14] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … & Schafer, B.

(2018). AI4People An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

[15] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*. https://arxiv.org/abs/1803.09010

[16] Lum, K., & Isaac, W. (2016). To predict and serve? *Significance, 13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

[17] Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679

[18] Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

[19] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

[20] Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

[21] Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481. https://doi.org/10.1145/3351095.3372828

[22] Raji, I. D., Smart, A., White, R. N., & Mitchell, M. (2022). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 33–44.

[23] Akinbolajo, O. (2024). The role of technology in optimizing supply chain efficiency in the American manufacturing sector. International Journal of Humanities Social Science and Management (IJHSSM), 4(2), 530–539.

[24] Chidozie et al. (2025). Quantum Computing and its Impact on Cryptography: The Future of Secure Communications and Post-Quantum Cryptography. 3. 10.5281/zenodo.15148534.

[25] Egbedion Grace et al. (2025). Securing Internet of Things (IoT) ecosystems: Addressing scalability, authentication, and privacy challenges. World Journal of Advanced Research and Reviews. 523-534. 10.30574/wjarr.2025.26.1.0999.

[26] Richardson, R., Schultz, J. M., & 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*(2), 192–233.

[27] Smith, E. (2020). UK government's exam results algorithm under fire for bias against poor students. *The Guardian*. https://www.theguardian.com/education/2020/aug/13/uk-exam-results-algorithm-criticized-for-unfairness

[28] West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race, and power in AI. *AI Now Institute*. https://ainowinstitute.org/discriminatingsystems.pdf