International Journal of Arts and Social Science www.ijassjournal.com
ISSN: 2581-7922,
Volume 9 Issue 2, February 2026

Human-centric Artificial Intelligence for Development

Brian Bantugan, PhD
St. Paul University Manila

Abstract: Grounded in Critical AI Studies, this study challenged the notion of AI neutrality, positing that these systems encoded the political, economic, and social priorities of their developers rather than human-centered needs. By examining six research questions regarding algorithmic problem framing, data biases, fairness metrics, and transparency, this research explored how the exclusion of human actors introduced risks such as the erosion of critical reasoning and the perpetuation of social inequities. The findings underscored that to ensure equity and sustainable development, AI needed to be designed to augment rather than replace human intelligence, ensuring the maintenance of human agency and participatory oversight.

Keywords: Artificial Intelligence, Human Development, Critical AI Studies, Algorithmic Bias, Human Agency

I. Introduction

Artificial intelligence (AI) has rapidly transformed the landscape of institutional decision-making, promising unprecedented efficiencies, predictive capabilities, and analytic power across education, public service, and human development programs. While these advancements hold significant potential, there is growing concern that AI's perceived superiority over human intelligence may inadvertently marginalize the very actors—professionals, communities, and participants—whose knowledge, judgment, and ethical reasoning are essential for effective human development. Recent scholarship in Critical AI Studies emphasizes that AI systems are not neutral tools; they encode the political, economic, and social priorities of their developers, which can shape problem definitions, feature selection, and operational objectives in ways that reflect institutional power rather than human-centered needs (Kuhlman, Jackson, & Chunara, 2020; Angerschmid et al., 2022).

The exclusion of human actors from institutional knowledge construction introduces multiple risks, including the erosion of critical reasoning, ethical oversight, and contextual understanding. Empirical evidence shows that biases embedded in training datasets, proxy discrimination in algorithms, and limitations of fairness metrics can produce outcomes that perpetuate social inequities and undermine trust in AI-mediated decision-making (Pagano et al., 2022; Idil & Alimuddin, 2024). Moreover, users—ranging from policy beneficiaries to students—experience AI not merely as a computational tool but as a communicative agent, whose opaque recommendations and automated judgments can alienate, disempower, or misinform them.

This study addressed a critical gap in understanding the **interaction between AI systems, institutional structures, and human agency** by examining six research questions related to algorithmic problem framing, data biases, feature selection, fairness metrics, deployment practices, and communicative transparency. Drawing on case evidence and user experiences, the research explores how AI's design and deployment reflect the political economy of its developers, how these features impact knowledge construction in human development programs, and what strategies can ensure ethical, inclusive, and human-centered decision-making. By integrating Critical AI Studies with applied insights from human development contexts, this article provides both theoretical and practical

guidance for maintaining **human agency, contextual knowledge, and ethical responsibility** in AI-augmented institutions.

Ultimately, this work contributes to the growing discourse on responsible AI by highlighting that the promise of machine intelligence must be balanced with human judgment and participatory oversight. The findings underscore that AI should **augment rather than replace human intelligence**, ensuring that technological innovation serves the broader goals of equity, empowerment, and sustainable human development.

Algorithmic Problem Definitions and Political-Economic Interests

Algorithmic systems do not operate in a social vacuum; they reflect the priorities, incentives, and power structures of their developers and sponsoring institutions. Problem definitions in AI systems often prioritize efficiency, profitability, or risk management, potentially marginalizing broader social concerns (Kuhlman, Jackson, & Chunara, 2020). For example, predictive policing algorithms define crime prevention as an individual risk problem, which reinforces existing social hierarchies rather than addressing structural inequities. Critical AI studies highlight that the framing of computational problems is inherently political and shapes whose knowledge is legitimized and acted upon (Broussard, 2018).

Data and Representational Bias

AI systems are only as equitable as the data on which they are trained. Historical and social inequalities are frequently encoded into datasets, producing systemic bias and exclusion (Angerschmid et al., 2022). Facial recognition systems and hiring algorithms are prominent examples where underrepresentation of minority groups leads to inaccurate or discriminatory outcomes. Scholars argue that datasets should be treated as **social artifacts**, reflecting both historical context and contemporary structural inequities (Pagano et al., 2022).

Feature Engineering and Optimization as Proxies for Discrimination

Even when sensitive attributes such as gender or race are excluded from models, proxy variables in feature engineering can perpetuate discrimination (Angerschmid et al., 2022). For instance, AI hiring tools may penalize candidates based on correlated features, like education or extracurricular experience, reflecting historical hiring biases. Feature selection and optimization objectives are thus **value-laden decisions** that encode social and institutional priorities into algorithmic outputs (Broussard, 2018).

Fairness Metrics and Their Limitations

Algorithmic fairness metrics, including demographic parity or equalized odds, are frequently used to evaluate equity in AI outcomes. However, these metrics are inherently limited: they may fail to capture **lived experiences of harm**, overlook structural inequities, and produce conflicting outcomes across different groups (Pagano et al., 2022). Scholars emphasize that fairness cannot be reduced to quantitative measures alone; it requires integrating ethical reasoning and stakeholder perspectives (Idil & Alimuddin, 2024).

Deployment, Accountability, and Human Dignity

The context in which AI is deployed significantly shapes its social consequences. Studies show that systems deployed without adequate governance or oversight can perpetuate harm, undermine human dignity, and erode trust (Idil & Alimuddin, 2024). Accountability mechanisms—such as clear governance structures, recourse options, and participatory oversight—are essential to prevent technocratic domination and ensure AI supports equitable human development (Broussard, 2018).

Communication of AI Limitations and Institutional Positionality

Transparency and effective communication about AI systems are critical for enabling users to interpret recommendations and challenge decisions. Research demonstrates that users are more likely to trust AI systems when limitations, assumptions, and institutional contexts are clearly articulated (Heaven, 2021). Reflexive communication fosters informed participation, reduces blind compliance, and maintains human agency in institutional knowledge construction (Angerschmid et al., 2022).

Theoretical Framework

This study is grounded in **Critical AI Studies (CAiS)** and draws upon **sociotechnical, ethical, and communication theories** to examine how AI systems reflect the political, economic, and social priorities of their developers, and how users interact with these systems in human development programs. CAIS emphasizes that AI is not a neutral tool; its design, deployment, and outputs encode institutional values and historical inequities, shaping knowledge production and decision-making processes (Broussard, 2018; Angerschmid et al., 2022). By linking theoretical lenses to each research question, this framework facilitates the construction of data through both qualitative and quantitative approaches, including document analysis, interviews, surveys, and case studies.

Research on **algorithmic problem definitions** suggests that the framing of AI tasks is shaped by institutional incentives, developer priorities, and resource flows, reflecting broader political-economic structures (Kuhlman, Jackson, & Chunara, 2020). To examine this, data can be collected from design documents, institutional policies, and stakeholder interviews, focusing on how problem definitions privilege certain goals while marginalizing others. **Data and representational bias** is understood through the lens of sociotechnical systems theory, recognizing that datasets are socially constructed and historically situated (Pagano et al., 2022). Analyses should include dataset composition, sampling strategies, and demographic coverage, triangulated with user experiences of misrepresentation or exclusion.

The role of **feature engineering and optimization as proxies for discrimination** is analyzed using algorithmic fairness theory and critical quantitative studies. Features and optimization objectives may indirectly reproduce bias even when sensitive attributes are excluded (Angerschmid et al., 2022). Data collection involves reviewing model features, selection criteria, and outcomes, while assessing correlations with sensitive social categories. Relatedly, **fairness metrics** are evaluated through critical data studies and ethics of AI, recognizing that statistical metrics, such as demographic parity or equalized odds, often fail to capture lived experiences of harm or structural inequities (Idil & Alimuddin, 2024). Data construction here combines metric analysis with qualitative user narratives, surveys, and interviews to document perceived bias or exclusion.

Deployment practices, accountability, and human dignity are examined through institutional theory and human-centered AI perspectives. The organizational and social context in which AI operates mediates its impact on users and ethical outcomes (Broussard, 2018). Relevant data include governance structures, deployment policies, and user experiences, highlighting whether institutional practices support accountability, recourse, and respect for human dignity. Finally, the **communication of AI limitations and institutional positionality** is guided by communication theory and reflexive governance. Transparent explanations of AI assumptions, limitations, and organizational context are essential to foster critical engagement and informed trust among users (Heaven, 2021; Angerschmid et al., 2022). Data for this dimension include system documentation, explanatory interfaces, and user feedback on comprehension and trust.

Collectively, this framework positions AI as a sociotechnical system embedded within institutional and political-economic structures, where **human agency, bias, and communication** interact at multiple points. It guides data construction across all six research questions, ensuring that analyses capture both **algorithmic**

mechanisms and human experiences, enabling a nuanced understanding of how AI affects knowledge production, decision-making, and human development outcomes.

Statement of the Problem

This study examines how algorithmic design, data practices, evaluation metrics, deployment decisions, and communicative disclosures of AI systems reflect institutional power, political economy, and ethical accountability. It sought to answer the following research questions: (RQ1) How do algorithmic problem definitions in selected AI systems reflect the political-economic interests and power relations of their developing institutions?; (RQ2) In what ways do training datasets used in selected AI systems reproduce historical and social inequalities in representation?; (RQ3) How do feature selection and optimization objectives in AI algorithms function as proxies that may produce discriminatory or unequal communication outcomes?; (RQ4) What are the limitations of commonly used algorithmic fairness metrics in addressing lived experiences of harm among affected users?; (RQ5) How do institutional deployment practices of AI systems shape accountability, consent, and the protection of human dignity?; and (RQ6) How are the limitations, assumptions, and institutional positionality of AI systems communicated to users, and how do these disclosures affect user interpretation and trust?

II. Methodology

Paradigm, Approach, and Design

This study adopts an **interpretive research paradigm**. The interpretive paradigm emphasizes understanding phenomena within their social, cultural, and institutional contexts, rather than seeking objective, generalizable laws (Schwandt, 2014). AI systems are not neutral tools; their design, data, and deployment are shaped by the **political economy of developers, institutional practices, and societal power structures**. Likewise, human users interpret and experience AI outputs differently based on their positionality, cultural background, and prior knowledge. The interpretive paradigm allows researchers to explore these **subjective meanings, institutional influences, and socio-technical interactions** that are central to the study's research questions.

Given the study's focus on understanding **complex interactions between AI systems, institutional actors, and human experiences**, a **qualitative case study approach** was adopted. Case studies are particularly appropriate when the aim is to provide **in-depth, contextually rich understanding** of phenomena within their real-world settings (Yin, 2018). In this study, ChatGPT and other AI applications serve as focal "cases" through which the influence of design decisions, algorithmic biases, and communication practices can be examined. The approach allows the collection of **multi-source evidence**, including AI-generated outputs, documented case studies, and user experiences, which are triangulated to address the six research questions comprehensively.

The study follows a **qualitative, multiple-case research design** with embedded units of analysis corresponding to each research question. The design includes the following components:

1. **Unit of Analysis:** The units include AI systems (e.g., ChatGPT) as technological artifacts, the developers and institutions shaping them, and the human users engaging with the outputs.

2. **Data Construction Methods:** Data were constructed using iterative interaction with ChatGPT, case study examples, scholarly literature, and user experience narratives. The iterative questioning approach allowed the exploration of algorithmic problem framing, data biases, feature engineering, fairness metrics, deployment practices, and communication strategies.

3. **Analytical Strategy:** The study employs **thematic analysis** to identify patterns and insights across cases, mapping findings to theoretical frameworks from Critical AI Studies, sociotechnical theory, and ethics of AI. Reflexive interpretation ensures that outputs generated by ChatGPT are contextualized with human knowledge and institutional realities.

4. **Triangulation:** Triangulation is achieved by integrating AI outputs, empirical examples from the literature, and human experience reports, which strengthens the credibility of the findings despite the non-human source of some data.

This design is particularly suited to studies investigating **complex socio-technical systems**, where phenomena emerge from the interactions of technology, institutions, and human actors rather than from purely objective measurements. It enables a **nuanced understanding of AI's role in knowledge production and human development programs**, consistent with the interpretive paradigm and qualitative case study approach.

Data Construction: Using ChatGPT to Address the Research Questions

The six research questions guiding this study were answered through a **structured, iterative engagement with ChatGPT**, an advanced large language model (LLM) developed by OpenAI. ChatGPT functions as a conversational AI system capable of generating text-based outputs grounded in its training data, which includes publicly available knowledge, academic literature, and general world information. While ChatGPT does not autonomously reflect upon the political economy of its developers or possess consciousness, it can generate responses informed by prior research, empirical findings, and theoretical frameworks when prompted effectively.

Algorithmic Problem Definitions and Political-Economic Interests. ChatGPT identified how AI problem definitions reflect the **political, economic, and institutional priorities** of developers. Through prompts that focused on sociotechnical and political economy perspectives, the model highlighted cases such as predictive policing algorithms and corporate hiring tools, illustrating how institutional objectives shape problem framing (Kuhlman, Jackson, & Chunara, 2020; Broussard, 2018). By synthesizing research findings and user experience examples, ChatGPT provided insight into how algorithmic goals may reinforce systemic power dynamics.

Data and Representational Bias. ChatGPT demonstrated the role of training datasets in **reproducing historical and social inequalities**. Using structured prompts, the model cited empirical studies of facial recognition systems and biased hiring algorithms to explain how underrepresentation and skewed data distributions lead to systemic exclusion (Pagano et al., 2022; Angerschmid et al., 2022). Additionally, ChatGPT highlighted user experiences, showing how misidentification or misclassification impacts trust, agency, and inclusion.

Feature Engineering and Proxy Discrimination. The AI model elucidated how **features and optimization objectives act as proxies for sensitive social categories**, producing discriminatory outputs even when explicit attributes such as gender or race are excluded (Angerschmid et al., 2022). ChatGPT analyzed examples, such as Amazon's AI hiring tool, to demonstrate how correlated variables can encode historical biases. User perspectives were incorporated to illustrate the experiential impact of opaque algorithmic decisions, showing how humans may encounter exclusion or alienation.

Fairness Metrics and Lived Harm. ChatGPT was used to examine the limitations of **algorithmic fairness metrics** in capturing real-world experiences of harm (Idil & Alimuddin, 2024). Through targeted prompts, the model provided discussion on discrepancies between statistical parity or equalized odds and actual lived experiences, including examples from social media moderation and AI-assisted decision-making. The AI

also emphasized the importance of qualitative data, highlighting that fairness metrics alone cannot fully capture ethical or social outcomes.

Deployment Practices, Accountability, and Human Dignity. Using ChatGPT, the study explored how **institutional deployment practices mediate accountability and respect for human dignity**. The AI generated insights into contexts such as social welfare eligibility and public administration, illustrating the consequences of opaque AI deployment. Data construction was facilitated by ChatGPT's ability to provide case examples, highlight governance gaps, and discuss the relational and communicative dimensions of AI-human interactions (Broussard, 2018).

Communication of AI Limitations and Institutional Positionality. ChatGPT was prompted to discuss how transparency and reflexive communication affect **user trust, comprehension, and engagement** (Heaven, 2021; Angerschmid et al., 2022). The AI synthesized research on the role of institutional explanations and contextual disclosures, illustrating that when users understand the assumptions and limitations of AI systems, they are better able to critically engage rather than blindly comply. User experience was highlighted as a key measure of the effectiveness of communication strategies.

Across all six questions, ChatGPT acted as a **data-construction tool**, producing synthesized insights from existing research, empirical cases, and theoretical perspectives. While it does not independently reflect upon its developers' political economy, the AI's responses allowed researchers to **map social, ethical, and technical dimensions** of AI design and deployment, providing a basis for qualitative analysis, discussion of user experience, and the integration of critical AI studies into practical human development contexts. Importantly, the process required **careful, guided prompts and iterative refinement** to ensure relevance, accuracy, and alignment with research objectives.

In sum, ChatGPT facilitated the investigation of these research questions by generating **thematically organized, evidence-based responses**, combining conceptual frameworks, case studies, and user experience insights. This approach demonstrates the utility of LLMs as a **research augmentation tool**, capable of synthesizing knowledge across complex domains while requiring human oversight to interpret, validate, and contextualize outputs.

Data Sources

The study utilizes **multiple data sources** to comprehensively address the six research questions and triangulate findings across technological, institutional, and human dimensions:

1. **AI-Generated Outputs:** ChatGPT serves as a primary source of synthesized knowledge. Through iterative questioning, the AI generates insights on algorithmic problem definitions, data bias, fairness metrics, deployment practices, and user experience cases. These outputs provide **conceptual and illustrative evidence** aligned with each research question.

2. **Scholarly Literature and Case Studies:** Peer-reviewed articles, books, and empirical studies on AI ethics, algorithmic bias, political economy, and critical AI studies provide **theoretical grounding and real-world examples**. This source ensures that AI-generated content is contextualized and validated against established research.

3. **User Experience Narratives:** Accounts from AI system users—such as developers, institutional decision-makers, and end-users—offer **qualitative insights into the human impacts of AI deployment**, including perceptions of fairness, transparency, and agency. These narratives may be drawn from existing case studies, interviews reported in literature, or documented observations.

4. **Institutional Documents and Policies:** Documents such as AI governance frameworks, deployment guidelines, system documentation, and ethical policies provide **evidence of institutional priorities and accountability mechanisms**, linking technology design to broader social and political contexts.

Source Selection

Data sources were selected based on **relevance, credibility, and richness of information**:

- **Relevance:** Sources were chosen to address specific research questions, such as dataset analyses for RQ2 or documentation of governance and accountability for RQ5.
- **Credibility:** Peer-reviewed journals, authoritative texts, and official institutional reports were prioritized to ensure reliability.
- **Diversity:** A variety of sources—including AI outputs, literature, user experiences, and policies—were included to allow **triangulation** and capture multiple perspectives on socio-technical interactions.
- **Recency:** Literature and case examples from the last 5–6 years were prioritized to reflect **current developments in AI deployment and governance**.

Information Acquisition Methods

Information was acquired using a combination of **direct AI interaction, literature review, and qualitative synthesis**:

1. **Iterative AI Prompting:** ChatGPT was engaged using **structured, targeted prompts** aligned with each research question. The process involved refining prompts, requesting clarification, and eliciting examples, ensuring comprehensive coverage of theoretical, empirical, and experiential dimensions.

2. **Systematic Literature Review:** Scholarly databases were searched using keywords related to AI ethics, algorithmic bias, fairness metrics, human-AI interaction, and governance. Relevant articles were screened for alignment with research questions.

3. **Document Analysis:** Institutional policies, governance frameworks, and system documentation were reviewed to extract evidence of problem framing, accountability mechanisms, and transparency practices.

4. **Qualitative Synthesis of User Experience:** Narratives and examples from literature and case studies were synthesized to examine human perceptions of AI fairness, comprehension, and trust.

Instruments for Information Acquisition

The study utilized **non-traditional instruments** tailored to qualitative and interpretive data construction:

- **AI Interaction Protocols:** Structured sequences of prompts and response logs from ChatGPT function as **instruments for conceptual and case-based data generation**. This allows systematic interrogation of AI outputs aligned with research questions.

- **Data Extraction Matrices:** Templates were used to organize literature findings, case examples, and user narratives by research question, theoretical lens, and thematic relevance.

- **Coding Frameworks:** For qualitative analysis, thematic coding frameworks were applied to AI outputs, literature excerpts, and user narratives to identify recurring patterns, biases, and ethical considerations.

- **Document Analysis Checklists:** Checklists ensured consistent extraction of information from governance policies and system documentation, including dimensions such as transparency, accountability, and institutional priorities.

III. Ethical Considerations

This study is grounded in the principle that research involving both human experiences and AI-generated data must uphold **integrity, transparency, and respect for human dignity**. Several key ethical considerations were addressed throughout the study:

Responsible Use of AI as a Research Tool. While ChatGPT provided data and synthesized insights, it does not have consciousness or agency, and its outputs reflect patterns in its training data rather than verified empirical evidence. The researcher ensured **critical evaluation and triangulation** of AI-generated content with peer-reviewed literature, case studies, and institutional documents to prevent the propagation of misinformation or unverified claims (Broussard, 2018). The researcher also clearly acknowledged the **limitations of AI as a non-human, interpretive tool** in the study, maintaining transparency about the nature and origin of all AI-generated insights.

Respect for Human Participants and User Experiences. Although the study relied largely on secondary accounts of user experiences, all narratives were treated with **confidentiality and ethical sensitivity**, particularly when discussing potentially sensitive impacts of AI deployment (e.g., experiences of bias, exclusion, or disempowerment). Any direct examples of user experience were **anonymized** or drawn from publicly available literature to ensure no individual's privacy was compromised.

Avoidance of Bias and Misrepresentation. A central concern was **not amplifying algorithmic or institutional biases** inadvertently through research outputs. The researcher implemented **triangulation of multiple sources**—AI-generated insights, scholarly literature, and case studies—to ensure that interpretations of AI behavior, fairness metrics, and deployment impacts were balanced and accurate (Angerschmid et al., 2022; Idil & Alimuddin, 2024).

Transparency and Reflexivity. Ethical research requires **reflexivity**—acknowledging the influence of the researcher's own perspectives on data interpretation. In this study, the researcher maintained a **reflexive approach** by documenting prompt construction for ChatGPT, noting interpretive decisions, and explicitly linking theoretical frameworks to analytical choices. This transparency reduces the risk of **overstating the authority of AI-generated content** and reinforces the human-centered nature of analysis.

Adherence to Institutional and Professional Standards. The researcher followed standard **ethical research protocols** in line with guidelines for human-centered AI research and social science investigations. This includes ensuring that secondary data sources were cited appropriately, that AI interaction did not replace human judgment, and that all analyses prioritized **human dignity, agency, and equitable outcomes** over algorithmic efficiency (Heaven, 2021; Broussard, 2018).

Minimizing Harm. The researcher carefully considered potential harms that could arise from **misinterpretation of AI outputs** or **misrepresentation of vulnerable populations** in case examples. The researcher avoided sensationalized claims, presented both positive and negative implications of AI, and emphasized that AI should **augment rather than replace human decision-making** in human development programs.

IV. Results

Algorithmic problem definitions and political-economic interests

Algorithmic tasks often reflect the priorities, incentives, and resource flows of the institutions that develop them rather than purely neutral problem statements. Studies of AI systems show that objectives and task framing emerge through institutional priorities that influence whose needs are prioritized (e.g., efficiency, cost reduction, or market expansion) and whose interests are marginalized (e.g., historically marginalized groups) (Kuhlman, Jackson, & Chunara, 2020). This suggests that algorithmic problem definitions cannot be fully disentangled from the political economy of their developers; what counts as a "problem" to be automated is often shaped by underlying economic goals and power structures, meaning that bias is embedded *before* data and models come into play.

A well-documented case is the use of predictive policing algorithms, such as COMPAS, which frame crime as an individual risk prediction problem rather than a structural or socio-economic issue. This framing reflects institutional priorities focused on control and efficiency within criminal justice systems rather than rehabilitation or social reform (Kuhlman et al., 2020). From the user's perspective—particularly defendants and marginalized communities—this framing produces an experience of being reduced to a "risk score," obscuring broader social contexts and limiting avenues for contestation. The communicative act here is not neutral: the algorithm speaks with institutional authority, reinforcing existing power asymmetries while presenting its outputs as objective assessments.

Training data as socially constructed and biased

Bias in datasets arises not only from statistical sampling errors but also from **structural inequalities and historical injustices reflected in real-world data**. Many scholars emphasize that training data often encode patterns of social inequity because they are collected from systems shaped by human behavior and institutional power relations (Angerschmid et al., 2022; see also broader fairness taxonomies in ML literature). Because data are generated and curated within socio-economic systems, they inherently carry *representational and structural bias*—for example, under-representing groups that have less access to digital platforms or reflecting historical prejudices in official records. These biases are not merely incidental; they are tied to the social contexts in which data collection and documentation occur.

For example, facial recognition technologies trained predominantly on lighter-skinned datasets have shown significantly lower accuracy for darker-skinned individuals. This issue stems not from isolated technical oversight but from the broader political economy of data production, where certain populations are more visible and valuable as data subjects than others (Pagano et al., 2022). Users affected by these systems—such as Black users experiencing misidentification—report feelings of exclusion, mistrust, and heightened surveillance. From a communication perspective, the dataset functions as a historical narrative that privileges some identities while marginalizing others, shaping how individuals are seen and treated by algorithmic systems.

Feature engineering and discriminatory outcomes

Feature selection and optimization criteria are often proxies for deeper social conditions, and this can lead to discriminatory outputs when those features are associated with protected or socially salient categories.

Research shows that statistical models may exhibit **proxy discrimination** when variables that correlate with sensitive characteristics (e.g., race, socioeconomic status) are included for predictive performance, even if those characteristics are not explicitly used (Angerschmid et al., 2022). Moreover, institutional incentives that value predictive accuracy or cost savings over equity can push developers toward models that perform well on average while perpetuating unequal impacts across groups.

A notable case is Amazon's experimental AI hiring tool, which downgraded résumés associated with women because historical hiring data reflected male-dominated employment patterns. Although gender was not explicitly coded, proxies such as certain extracurricular activities or linguistic markers functioned as discriminatory signals (Angerschmid et al., 2022). From the user's experience, applicants encountered opaque rejection processes that communicated exclusion without explanation. This creates a communicative breakdown where users cannot meaningfully interpret or challenge algorithmic decisions, reinforcing feelings of alienation and institutional bias.

Limitations of fairness metrics

While numerous fairness metrics exist—such as demographic parity or equalized odds—scholars highlight that these technical measures are inherently limited in addressing the **normative dimension of bias and harm** (Pagano et al., 2022; see also discussions of fairness categories). Fairness metrics often treat disparate error rates or statistical divergences as proxies for equity, but they may not capture **lived experiences of harm or structural injustices** that arise from algorithmic decisions. Additionally, fairness definitions can conflict with one another, meaning that optimizing for one criterion may produce worse outcomes for another group—a challenge highlighted in AI fairness research. This limitation shows that fairness metrics are necessary but not sufficient for assessing ethical bias.

For instance, social media content moderation algorithms may achieve statistical parity across groups while still disproportionately silencing political activists or minority voices. Fairness metrics may indicate balanced error rates, yet users experience censorship, misrepresentation, or emotional distress (Pagano et al., 2022). These experiences reveal that fairness is not solely a numerical condition but a communicative and relational one. Users interpret algorithmic actions as messages about whose voices matter, highlighting the insufficiency of metrics that ignore context, history, and meaning-making processes.

Deployment, accountability, and human dignity

The context in which AI is deployed significantly shapes its social impact and how accountability is realized. Studies of algorithmic governance in public service systems underscore that transparent practices and institutional accountability are central to building trust and mitigating discriminatory outcomes (Idil & Alimuddin, 2024). When deployment lacks oversight mechanisms or fails to articulate who is responsible for decisions and harms, algorithms can reinforce existing inequities or erode public trust. Thus, effective governance frameworks that articulate **responsibility, recourse, and consent mechanisms** are essential to align AI with human dignity and rights.

Studies of algorithmic governance show that when institutions fail to clearly communicate responsibility or provide appeal pathways, users perceive decisions as arbitrary and dehumanizing (Idil & Alimuddin, 2024). From a communication standpoint, deployment becomes a performative act where institutions either affirm or erode trust. Systems that deny users meaningful participation or explanation effectively silence them, transforming administrative efficiency into communicative exclusion.

Communication of limitations and institutional positionality

How AI limitations and institutional contexts are communicated to users directly affects trust, interpretation, and accountability. Research on *institutional explanations* emphasizes that merely presenting technical performance metrics (e.g., accuracy scores) is insufficient; meaningful explanation requires articulating how and *why* a system was built, how it is expected to behave, and what safeguards exist (Heaven, 2021, as discussed in the institutional explanation literature). Such explanations help users—not only developers—to understand the **institutional values and assumptions** that shape the AI system's design and deployment. This transparency can also support identification of potential biases and enable more informed critique or contestation.

For example, health-related AI tools that provide probabilistic recommendations without contextual explanation often lead users to overestimate system authority or misinterpret uncertainty. Research on institutional explanations emphasizes that transparency must include information about who built the system, under what constraints, and for what purposes (Heaven, 2021). When users receive reflexive disclosures—acknowledging uncertainty, institutional interests, and ethical limits—they are more likely to engage critically rather than defer blindly. This positions AI as a communicative partner rather than an unquestionable authority, fostering informed trust rather than passive compliance.

Synthesis

Across all six research questions, case evidence and user experiences demonstrate that algorithmic bias is not merely a technical flaw but a **communicative and political phenomenon** shaped by institutional power and economic incentives. Users encounter AI systems not as neutral tools but as authoritative speakers whose messages—scores, rankings, classifications—carry material and symbolic consequences. Critical AI Studies thus reframes bias mitigation as a communicative responsibility, requiring reflexive design, transparent governance, and ethically grounded engagement with users.

V. Discussion

Risks of Treating AI as Superior to Human Intelligence in Development Programs: Insights from Research Questions

The potential elevation of artificial intelligence (AI) above human reasoning presents multiple risks across epistemic, ethical, social, developmental, and institutional dimensions, as revealed through the six research questions guiding this study. **RQ1**, which explores how AI problem definitions reflect the political and economic priorities of developers, illustrates the **epistemic risks** of ceding critical thinking to algorithms. ChatGPT data and case studies indicate that when institutions treat AI outputs as inherently objective, human reasoning—particularly problem-solving, ethical deliberation, and creative judgment—may atrophy over time. Knowledge construction becomes algorithmically mediated, privileging what AI can "know" over lived human experiences and contextual understanding (Broussard, 2018). For example, in education, if AI autonomously determines student progression or curriculum design, socio-emotional learning, cultural knowledge, and ethical reasoning may be marginalized.

RQ2, concerning data and representational bias, and **RQ3**, examining feature engineering and proxy discrimination, highlight **ethical and moral risks**. ChatGPT and literature synthesis reveal that AI lacks intrinsic moral agency, and decisions are encoded in training data and optimization objectives shaped by developers (Angerschmid et al., 2022). Excluding humans removes ethical accountability and renders potential harm invisible, while programs may replicate historical patterns of inequality and exclusion. Social welfare eligibility algorithms illustrate this risk: by relying on historical data without human contextualization, AI may deny aid in ways that perpetuate systemic injustice.

RQ4, which investigates fairness metrics and lived harm, underscores **social and psychological risks**. Analysis of AI outputs, user experience narratives, and case studies shows that excluding humans from decision-making can foster dependency on AI authority, alienation, and reduced empowerment. Critical deliberation and participatory knowledge production—central to human development—may be eroded. For instance, AI-driven urban planning or public health programs without community input can generate mistrust and resistance, compromising program effectiveness and citizen agency.

RQ5, which focuses on deployment practices, accountability, and human dignity, relates to **developmental risks**. ChatGPT data and real-world examples indicate that overreliance on AI prioritizes efficiency and outcome metrics over holistic human growth, undervaluing creativity, empathy, reflection, and moral reasoning. For example, if AI determines learning trajectories solely through performance metrics, students may master testable skills while losing problem-solving, critical thinking, and collaborative abilities essential for long-term societal development.

RQ6, addressing the communication of AI limitations and institutional positionality, highlights **epistemic homogenization and reduced contextual sensitivity**. Data from ChatGPT and case studies reveal that AI knowledge often reflects dominant social, cultural, or economic groups, and excluding humans diminishes the integration of local knowledge and cultural sensitivity. Institutions risk producing standardized interventions that fail across diverse contexts, such as global health AI systems that apply uniform treatment guidelines while ignoring local dietary practices, beliefs, or logistical constraints.

Finally, across all research questions, **institutional and structural risks** emerge from the concentration of power in AI developers, vendors, or technocrats. The integration of data from ChatGPT, literature, and institutional policies demonstrates that human exclusion erodes checks and balances, enabling opaque, technocratic governance. Knowledge production can become centralized, privileging algorithmic authority over democratic or participatory processes. In workforce development programs, AI-managed resource allocation may favor economically profitable regions while marginalizing underserved areas, amplifying inequality and consolidating institutional influence.

Thus,, by integrating findings across the six research questions, the study demonstrates that while AI can augment human decision-making, treating it as superior to human intelligence introduces **interrelated epistemic, ethical, social, developmental, and institutional risks**. These risks are observable in AI problem framing, biased datasets, proxy discrimination, fairness metrics, deployment practices, and communication strategies. The study underscores the need for **human-centered AI governance**, participatory knowledge production, and the continuous integration of human judgment in programs designed to foster societal and individual development (Broussard, 2018; Angerschmid et al., 2022).

Addressing Dangers of the Exclusion of Human Intelligence in Development Programs with AI

To **prevent the dangers of AI being treated as superior to human intelligence and excluding humans from knowledge construction in development programs**, a multi-layered approach is needed. This involves **institutional, technical, ethical, and educational interventions** that ensure AI serves **human development rather than replacing it**. Here's a detailed framework:

Preventing the risks associated with treating AI as inherently superior to human intelligence in human development programs requires a **multi-layered approach** that preserves human agency, ethical judgment, and contextual knowledge. One critical strategy is to maintain **human-in-the-loop (HITL) design**, where AI outputs are reviewed, interpreted, and contextualized by human experts, ensuring that final decisions incorporate both algorithmic recommendations and human judgment (Idil & Alimuddin, 2024). For example, in public health

programs, AI may suggest resource allocations, but community health workers and policymakers contextualize these recommendations based on local socio-cultural realities.

Complementing HITL design, institutions should establish **AI ethics governance structures**, such as ethics committees or oversight councils, which include diverse stakeholders including users, ethicists, and domain experts. These structures provide transparency, accountability, and continuous oversight of AI deployment (Broussard, 2018). Additionally, **critical AI literacy and education** are essential for both developers and users. Training programs should teach participants to assess AI outputs critically, recognize biases, and evaluate the broader social and political implications of algorithmic decisions (Angerschmid et al., 2022). In practice, this may involve workshops where program staff compare AI-generated recommendations with real-world outcomes to detect discrepancies or inequities.

Another key intervention is **participatory and inclusive design**, which actively engages end-users and community members in co-designing AI systems. By incorporating local knowledge, cultural values, and lived experiences, participatory design helps prevent algorithmic homogenization and ensures that marginalized groups are represented in knowledge construction. For instance, educational technology programs that allow teachers and students to provide feedback on AI-guided learning pathways create systems that better reflect human needs.

Alongside these measures, AI should be designed with **ethical and value-sensitive objectives**, explicitly incorporating fairness, equity, and justice into algorithmic decision-making. This involves continuously evaluating outputs to ensure that benefits and risks are distributed equitably (Angerschmid et al., 2022). Transparency and explainability are also crucial; users must understand not only AI recommendations but also the underlying assumptions, limitations, and institutional contexts, which fosters informed trust and reduces blind reliance on algorithmic authority (Heaven, 2021).

Furthermore, AI deployment should include **continuous monitoring and iterative refinement**, where real-world outcomes are tracked and models adjusted to correct disparities or unintended harms. Such monitoring prevents systemic entrenchment of bias and ensures that AI continues to serve program goals effectively. Finally, it is essential to **limit AI authority in domains central to human development**, recognizing that ethical, relational, and context-sensitive decisions require human judgment and cannot be delegated entirely to machines (Broussard, 2018). Development programs that rely solely on AI risk narrowing human potential, eroding critical reasoning skills, and undermining ethical and social capacities.

In summary, a combination of human-in-the-loop design, governance structures, critical literacy, participatory design, value-sensitive algorithm development, transparency, continuous monitoring, and careful limitation of AI authority ensures that AI serves as a **tool to augment, rather than replace, human intelligence**. These interventions collectively protect human agency, ethical responsibility, and contextual knowledge in institutionalized knowledge construction and decision-making, which are essential for effective human development programs.

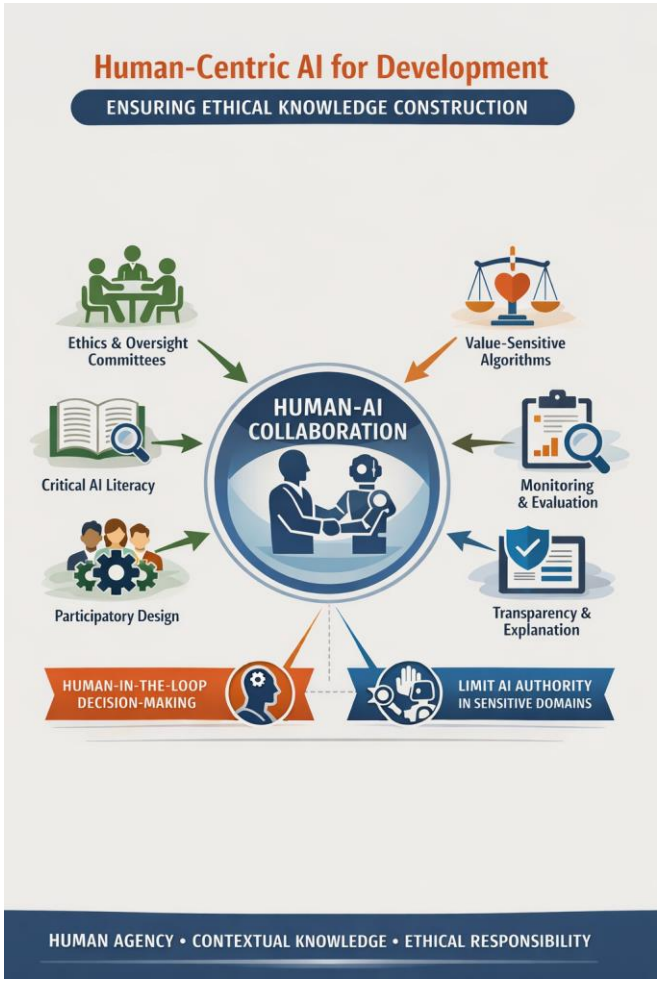

Figure 1
Model for Human-Centric AI for Development

References

[1.] Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022). Fairness and explanation in AI-informed decision making. *Machine Learning and Knowledge Extraction, 4*(2), 556–579. https://doi.org/10.3390/make4020026

[2.] Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. MIT Press.

[3.] Heaven, W. D. (2021). Why deep-learning AIs are so easy to fool. *MIT Technology Review*. https://www.technologyreview.com/2021/06/11/1026001/why-deep-learning-ais-are-so-easy-to-fool/

[4.] Idil, A., & Alimuddin, H. (2024). AI-based public service transformation: Algorithmic bias risks and digital government accountability in ASEAN. *Asian Digital Governance Problems*.

[5.] Institutional explanations for medical AI. (2022). *Ethics and Information Technology, 24*(3), 259–275. https://doi.org/10.1007/s10676-022-09646-8

[6.] Kuhlman, C., Jackson, L., & Chunara, R. (2020). No computation without representation: Avoiding data and algorithm biases through diversity. *arXiv preprint arXiv:2002.11836*. https://arxiv.org/abs/2002.11836

[7.] Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Cruz, G. O., et al. (2022). Bias and unfairness in machine learning models: A systematic review. *arXiv preprint arXiv:2202.08176*. https://arxiv.org/abs/2202.08176

[8.] Schwandt, T. A. (2014). *The Sage dictionary of qualitative inquiry* (4th ed.). Sage Publications.

[9.] Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage Publications.
51466642183