Decolonizing Artificial Intelligence: Unveiling Biases, Power Dynamics, and Colonial Continuities in AI Systems

Author: Areeba Kamran
Lahore, Punjab, Pakistan

ABSTRACT: Artificial intelligence (AI) holds immense potential for diverse applications, but ethical concerns arise due to its growing ability to learn, decide, and act in uncertain environments. As AI evolves, the concept of a potential "intelligence explosion" looms, with the emergence of a superintelligence surpassing human capabilities. However, defining ethical standards and decision-making criteria becomes complex, particularly in real-time conflicts where AI's decisions might override human authority. This paper explores the intricate interplay of biases in data collection, AI algorithms, and potential modern-day colonial implications.

Introduction.

The rapid advancement of artificial intelligence (AI) has ignited excitement and ethical concerns alike. As AI's capabilities expand, the potential for an "intelligence explosion" leading to superintelligence raises profound questions about ethical decision-making. The need to navigate the ethical intricacies of AI becomes even more pronounced in contexts of real-time conflict, where AI's autonomy might challenge human authority. While technical approaches aim to mitigate biases in AI systems, they often fall short due to the complex ideological underpinnings of data collection. This paper delves into the ethical challenges posed by biases in data and AI and their implications for contemporary society. It also explores the analogy of "colonial AI" and its relation to historical power dynamics.

Figure 1. Aníbal Quijano was a Peruvian sociologist known for his work on coloniality, modernity, and power dynamics, highlighting how colonial legacies persist in shaping societies and knowledge structures. Figure from https://www.udg.mx/sites/default/files/styles/actividad/public/anibal_quijano.jpg?itok=lYckNiTF

Literature Review.

Biases in data collection.

Data collection and curation are not neutral

processes but influenced by worldviews and ideologies, shaping AI systems' understanding of the world. Common technical approaches to address bias in machine learning focus on re-balancing data for fairness, but this is not enough. Critical studies of race, feminist theories, and social justice are proposed by (Susan et.al 2020) to understand power structures and ideologies embedded in the data. Integrating these perspectives prompts us to question who controls data collection, whose experiences are represented, and how societal inequalities manifest in the data. Developing democratic and inclusive ML requires considering data composition, resource capacity, and values. The fragmented discourse on each axis is problematic, hindering a true understanding of real-world dynamics (Maryam et.al, 2023).

Instead of solely fixing biases, the focus should be on designing datasets that intentionally represent desired values. Balancing underrepresented groups can lead to ethical concerns. Data regulation should be grounded in discrimination concepts and existing legal frameworks to ensure fairness and non-discrimination, providing legal protection to individuals (Susan et.al 2020). Collaborative ML shows promise for inclusivity. Eliminating bias from large language models is complex due to inherent biases in language and cultural norms. It involves separating useful patterns from ingrained biases, understanding diverse perspectives, defining fairness, and adapting to language changes (Emilio, 2023).

Data sharing is considered a means for inclusion and representation in ML models, but it raises concerns about data autonomy and agency, resembling colonial-era agreements. Trusting powerful entities with data use may lead to exploitation, limiting affected communities' control over their data. Alternative scenarios with more equitable power dynamics and collective ownership are possible but challenging to govern. Data-sharing alone does not guarantee fairness or benefits; model architecture, algorithmic design, and data nature also influence performance and fairness. Small datasets can lead to biased outcomes and compromised generalization (Maryam et.al, 2023).

Illusion of objectivity in data.

AI's attribution of objectivity and its performative nature make it challenging to address biases solely through algorithmic adjustments. The illusion of objectivity surrounding AI systems leads to unwarranted deference to their outputs, especially when expressed numerically. Numbers are perceived as unbiased and meaningful, but like AI systems, they are value-laden and context-dependent. AI bias differs from direct human bias in its apparent objectivity and performance. AI systems can create and enact their own social biases, making it more challenging to address their biases effectively (Dan L.Burk, 2022). Efforts to "debias" AI data assume that biases can be isolated and corrected discreetly, but this may not be the case. It is also presumed the existence of an "unbiased" baseline, which is challenging to define. The argument that AI will be more accurate than humans or will improve over time does not fully address the problem of social bias. Even if technical biases are removed, social bias can still persist in AI systems. Technical accuracy alone does not address the underlying societal issues that contribute to bias. The approach of correcting biased outcomes in AI systems may not be appropriate as machine-driven biases differ significantly from human biases. Machine-driven discrimination

might be more virulent or persistent, requiring different or more drastic solutions. Algorithmic bias differs from direct human bias in how it is perceived as more objective and neutral, which can be manipulated by system providers to advance their own agendas and maintain biased outcomes (Dan L.Burk, 2022).

Modern day colonialism.

"Critical race theorists have already observed that "AI" as a cultural phenomenon is coded white" (Dan L.Burk, 2022).

While formal colonial rule ended in various parts of the world in the 20th century, the enduring legacies of colonialism and racial divisions persisted. Post-colonial states faced new forms of imperialism that hindered their sovereignty through economic, legal, and epistemic means. The effects of colonialism persist, leading to the emergence of the broader notion of coloniality. Coloniality refers to the continuation of colonial characteristics and power dynamics in the present, influencing various aspects of culture, labor, intersubjectivity, and knowledge production. The concept of decoloniality, which has been interpreted differently in various local contexts. In South America, decoloniality emerged in the 1990s, emphasizing the interconnection between colonialism, modernity, and capitalism. Distinct from decolonisation, decoloniality aims to critique and reverse these subsequent imperialistic structures that emerged from the colonial era (Rachel, n.d). Decolonisation is seen as the restoration of land and life after the end of historical colonial periods, involving the undoing of territorial appropriation, exploitation of resources, and direct control of social structures.

Aníbal Quijano, Peruvian sociologist, and others define coloniality as the driving force behind modernity and capitalism. Quijano views decoloniality as the counterforce to coloniality, not merely a post-colonial deconstructive power. He draws from Foucault's ideas on power/resistance to envision 'de/coloniality' as the potential to unravel the coloniality of power, which perpetuates oppression and Eurocentric ways of knowing (Rachel, n.d). Quijano emphasizes the control over social structures in dimensions like authority, economy, gender, sexuality, and knowledge. Maldonado-Torres views coloniality as the perpetuation of hierarchies of race, gender, and geopolitics, employed as tools of colonial control. The concept is extended to the digital realm by Couldry and Mejias, who relate it to modern data relations that recreate a form of colonizing power (Mohamed et.al, 2020). Despite designers' intentions, AI systems inadvertently embed the prevailing values of those in power. These design choices and value homogenization lack sufficient documentation and transparency, leading to non-transparent control over users' rights. This reinforces existing power structures, creating an "us versus them" mindset. The rush towards large AI models further concentrates power, resulting in increased access inequality and a colonial AI state (Maryam et.al, 2023).

Breaking down decolonisation.

Decolonisation assumes two roles: territorial and structural. Territorial decolonisation involves dissolving colonial relations, while structural decolonisation seeks to undo colonial mechanisms of power, economics, language, culture, and thinking. It involves interrogating the origins and legitimacy of dominant knowledge, values, norms, and assumptions. Three views shed light on this decolonial knowledge landscape. A "decentring view" rejects imitating the West, emphasizing unique

identities and re-centering knowledge on global histories and diverse approaches. An "additive-inclusive view" incorporates new and alternative approaches alongside existing knowledge, fostering environments for diverse knowledge creation. An "engagement view" calls for critical examination of science from marginalized perspectives, questioning power dynamics, inclusivity, and unacknowledged assumptions (Mohamed et.al, 2020).

Three forms of discontent arise within the realm of AI according to (Rachel, n.d), who connects them to coloniality and the historical construction of race. First, the presence of racial and gender bias present in AI technologies exacerbates inequalities. Second, the commoditization of human experience within the data paradigm, leading to individuals being reduced to data points. The third form of discontent pertains to a geopolitical arms race among transatlantic nations for AI innovation. This, along with the second form, contributes to the concepts of 'data colonialism' and 'digital colonialism,' leading to the extraction and exploitation of personal data. The drive for AI dominance reflects hegemonic impulses and highlights the potential division between low-tech and advanced-tech states.

The metropole-periphery framework identifies centers of power and their peripheries, reflecting imbalances in authority and participation. Dependency theory extends this by connecting colonial histories to present-day underdevelopment and economic disparities. This framework helps analyze contemporary AI practices as features of colonial continuities, where technology corporations resemble metropoles of technological power. However, the limitation of the metropole-periphery framework, is the oversimplification of complex global dynamics. Which suggests incorporating diverse modes of decolonial thought, such as contrapuntal analysis, psychodynamic perspectives, economic analysis, and historical and literary criticism (Mohamed et.al, 2020). The language of decoloniality has led to new perspectives in AI ethics. Mohamed, Png, and Isaac propose "Decolonial AI" and advocate for dialogue between AI centers and peripheries to develop "intercultural ethics." This involves learning from peripherals through "reverse pedagogies" and embracing pluralism, pluriversal ethics, and local designs. Sabelo Mhlambi goes further by building an AI ethics framework based on the Nguni philosophy of Ubuntuism. He critiques Western rationality in shaping AI ethics and highlights Ubuntu's emphasis on relational personhood. Mhlambi's framework addresses challenges like surveillance capitalism and data colonialism (Rachel, n.d).

While some scholars mention the limitations and colonial aspects of universal ethics, there's a lack of in-depth exploration of the historical dominance of Eurocentric ethics and its connection to colonial practices. This raises concerns about uncritically incorporating decoloniality into AI ethics discussions, as it could inadvertently perpetuate the racial logics established by colonialism. (Rachel, n.d) highlights the potential risk of appropriating decoloniality as a metaphor and reproducing colonialism's racial frameworks.

Facial Recognition technologies

The use of facial recognition technologies in advanced biometric systems. These systems analyze facial characteristics and compare them to a database of facial images to determine various attributes like gender, race, age, and sexuality. The author highlights how these

technologies reinforce social hierarchies and racial inequalities by inferring social status from physical appearances. This approach replicates the historical logics of race and racism, where surface features were used to make judgments about cognitive abilities and behavior. (Rachel, n.d) also points out that contemporary facial recognition systems continue similar practices to infer intent, predict behavior, and even assess intelligence. Moreover, these systems can reintroduce race science by justifying their use under the pretense of security, market efficiency, and risk management.

Energy and environmental impact.

The mechanisations of recognition and interpretation, enabled by neural networks and sensors collecting big data, comes with a high energy cost (Lohmann, 2021). Contrary to the perception of a "dematerialized" economy, AI intensifies the same material violence against women, nature, colonies, people of color, and workers. Divisions of labor that originated in pre-industrial capitalism underpin both historical and modern industrial revolutions. Computer science and robotics have introduced new divisions of labor that support AI's massive volumes and speeds, relying on low-cost labor. The efficiency of unit computations has greatly increased, yet the overall system's violence and inefficiency remain.

(Lohnmann, 2021) suggests that AI's manipulation of entropy gradients and energy consumption is intertwined with evolving industrial technologies. The focus, however, should be on tracing and categorizing the different ways AI sets up new gradients between low and high entropy worldwide. These arrangements involve appropriations from workers and the environment and constitute instances of political violence. The commonalities between AI's effects and the 19th-century industrial revolution, including class conflict, capitalist competition, divisions of labor, and colonial projects to restore entropy slopes. It is suggested that AI generates "stupidity" similar to classical assembly lines, reinforcing mind/body dualisms central to contemporary society.

Generative Language Models

ChatGPT claims political neutrality but acknowledges the presence of biases in its training data. AI systems should avoid taking stances on issues where scientific evidence is inconclusive and be inclusive of all legal viewpoints. However, potential biases may arise from the large corpus of internet data used in training, which is dominated by influential Western institutions. (David, 2023) Raises concerns about AI systems claiming neutrality while displaying biases, as they can exert social control and shape human perceptions..
Generative language models learn from vast amounts of data, which allows them to generalize effectively. However, this generalization can perpetuate biases present in the training data, leading to unfair treatment and marginalization of certain groups. Biases can also emerge unexpectedly due to complex interactions within the model. Non-linear relationships mean that even small biases can have significant negative effects.

“Humans in the loop”.

AI's advantages in speed, efficiency, and accuracy are balanced by its limitations in contextual thinking and potential catastrophic failures in novel situations. This creates reluctance to grant the technology complete autonomy without human oversight (Mazzolin, 2020). While it's argued that removing human control from AI, especially AI-enabled

weaponry could improve the ethics of warfare, there are concerns. Human engagement in combat often leads to suboptimal decision-making due to stress, fatigue, and emotions, causing unintended collateral damage. AI, with its potential for objective decision-making, might outperform humans morally in similar situations, but the implications of fully autonomous weapons raise significant ethical and safety challenges.

The debate over AI-enabled weaponry revolves around the potential for more ethical conflict conduct due to AI's lack of emotional impulses and superior decision-making. However, concerns arise about reducing the threshold for war, prolonging conflicts, liability, and moral responsibility. The need to keep humans in the loop depends on AI's ability to discriminate between data sets and detect manipulation attempts, as current vulnerabilities exist that can be exploited by malicious actors. AI systems are confronted with decisions that hold life-and-death implications, be it in combat, healthcare, or on public roads. This raises concerns about the ethics of valuing human lives differently and relying on a "death algorithm" for independent decisions. Determining the basis for resource allocation in medical settings or distinguishing combatants from civilians poses significant challenges. The lack of a universally agreed-upon moral framework adds subjectivity, which current AI systems struggle to address. International governance bodies should carefully consider this issue while developing regulatory frameworks to tackle these ethical dilemmas (Mazzoln, 2020).

Challenge	Description
Inherent biases in language	Language mirrors society, harboring biases and stereotypes. Extracting valuable patterns from these ingrained biases poses difficulty due to their deep-rooted presence
Ambiguity of cultural norms	Encoding cultural norms in AI is intricate, demanding deep comprehension of diverse perspectives within communities and regions to make informed decisions about model representation.
Subjectivity of fairness	Defining "fairness" for AI models is complex due to subjective interpretations. Addressing bias necessitates contextual definitions aligned with diverse stakeholders and perspectives across applications.
Continuously evolving language and culture	AI models must adapt to evolving language and culture, incorporating new expressions, norms, and biases. Ongoing monitoring and adjustment are crucial to maintain unbiased

	performance.

Table 2. Challenges in addressing biases in Large Language Models. Table from https://arxiv.org/pdf/2304.03738.pdf

Discussion.

Biases in data collection significantly impact AI's understanding of the world. Addressing bias through data re-balancing is insufficient; critical perspectives, such as race and feminist theories, are needed to unravel the power structures embedded in data. Designing datasets that intentionally represent desired values rather than merely fixing biases is crucial. However, balancing underrepresented groups may introduce new ethical dilemmas. Data autonomy and agency in AI systems echo colonial-era agreements, underscoring the importance of data regulation rooted in discrimination concepts and legal frameworks to ensure fairness.

The illusion of objectivity surrounding AI systems complicates bias mitigation efforts. AI's attribution of objectivity and performative nature obscure biases, rendering algorithmic adjustments insufficient. Technical accuracy alone does not address the underlying societal issues contributing to bias. Efforts to "debias" AI data may not isolate and correct biases as presumed, and the potential for machine-driven biases to differ significantly from human biases underscores the complexity of the challenge.

Conclusions.

Technology by itself does not result in discrimination and injustice, but it can trigger individuals' perceptions, leading to either acceptance or rejection of biased information and technology (Nima and Maryam, 2021). This paper highlights that decolonizing AI involves critiquing how AI relies on and perpetuates colonial power structures and racial divisions. The discourse suggests that some of AI's inherent assumptions and paradoxes align with Western power dynamics, leading to questions about AI's potential to be truly decolonial. As AI advances globally, imagining a future with or without AI becomes essential, considering its potential to reinforce inequality or contribute to a more multifarious world. A shift in the approach to decolonization within the field of AI. Instead of asking how decolonial thought can be applied to AI, the author suggests considering how colonialism has shaped AI. Without adopting a decolonial perspective that critically situates AI within the historical context of colonialism and race, there's a risk of perpetuating the issues that decoloniality aims to challenge (Rachel, n.d).

The evolution of AI presents profound ethical challenges stemming from biases in data and algorithmic decision-making. The analogy of "colonial AI" serves as a cautionary framework, reminding us of the need to critically examine the values, assumptions, and power dynamics underpinning AI systems. As AI continues to shape various facets of contemporary life, a comprehensive understanding of biases and their implications is crucial to ensure a more equitable and just future.

Applications.

Additional scholarly work should focus on the challenges and risks of effectively overseeing complex AI systems. Societal impact analysis should include topics like algorithmic transparency and AI's effects on democracy. Collaboration with diverse communities is vital to address bias, evaluate, and create more equitable AI systems. Power dynamics must also be addressed to avoid one-sided control and

suppression of diverse perspectives. Prioritizing resource demands, accuracy, robustness, and defense against attacks are crucial considerations. Researchers must explore mitigation measures like AI system patching for software deficiencies and implementing "exit ramps" and "firebreaks" to align AI development with socially accepted standards as proposed by (Mazzolin, 2020).

Rigorous evaluation and mitigation techniques are needed to understand and address these biases, as randomized controlled trials may not be feasible for ethical reasons. The Reinforcement Learning with Human Feedback (RLHF) strategy is used to reduce biases by fine-tuning models using human demonstrations and preferences, but the potential for deliberate misalignment must also be considered (Emilio, 2023).

Limitations.

Some research topics required access to datasets or primary sources that are not readily available online. Some valuable scholarly resources are behind paywalls, restricting access for individuals without subscriptions or institutional affiliations. This limitation may hinder the inclusion of comprehensive and diverse perspectives in the research.

Acknowledgements.

I would like to acknowledge my mentor and the founder of RMS, Aatmi without whom I would have never been able to accomplish such a feat.

Author.

I am a high school student in Pakistan hoping to graduate in 2024 and pursue robotics and engineering in my future endeavors.

References.

Adams, R. (2021). Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews*, *46*(1–2), 176–197. https://doi.org/10.1080/03080188.2020.1840225

Birhane, A., & Guest, O. (2021). Towards decolonizing computational sciences. *Kvinder, Køn & Forskning*, (2), 60–73. https://doi.org/10.7146/kkf.v29i2.124899

Cazes , M., Franiatte, N., & Delmas , A. (2021). *Evaluation of the Sensitivity of Cognitive Biases in the Design of Artificial Intelligence.*

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., & Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and Healthcare. *Npj Digital Medicine*, *3*(1). https://doi.org/10.1038/s41746-020-0288-5

Engler, A., Michael Kearns, A. R., Manish Raghavan, S. B., MacCarthy, M., Regina Ta, D. M. W., & Chakravorti, B. (2023, June 27). *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*. Brookings. https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Ferrara, E. (2023). *Should ChatGPT Be Biased? Challenges and Risks of Bias in Large Language Models*.

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409.

https://doi.org/10.1080/0960085x.2021.1927212

L. Burk, D. (2022). *Racial Bias in Algorithmic IP*.

Leavy , S., O'Sullivan, B., & Siapera, E. (2020). *Data, Power and Bias in Artificial Intelligence*.

Lohnmann, L. (2021). *Heat, Colonialism and the Geography of Artificial Intelligence*.

Mazzolin , R. (2023). *Artificial Intelligence and Keeping Humans "in the Loop"*.

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial ai: Decolonial theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, *33*(4), 659–684. https://doi.org/10.1007/s13347-020-00405-8

Molamohammadi, M., Taik, A., Le Roux, N., & Farnadi, G. (2023). *Unraveling the Interconnected Axes of Heterogeneity in Machine Learning for Democratic and Inclusive Advancements*.

Panch, T., Mattie, H., & Atun, R. (2019). Artificial Intelligence and algorithmic bias: Implications for health systems. *Journal of Global Health*, *9*(2). https://doi.org/10.7189/jogh.09.020318

Rozado, D. (2023). The political biases of chatgpt. *Social Sciences*, *12*(3), 148. https://doi.org/10.3390/socsci12030148

Sen, P., & Ganguly, D. (2020). Towards socially responsible AI: Cognitive bias-aware multi-objective learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(03), 2685–2692. https://doi.org/10.1609/aaai.v34i03.565

Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine learning. *Proceedings of the 51st Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2018.668