A British med school was indicted for racial discrimination in 1988 after the UK Commission for Racial Equality discovered that its computer program used to select applicants for interviews was biased against women and non-European-named applicants (Lowry and Macpherson,1988). The computer program, however, had been developed to emulate human admissions decisions with an accuracy rate of between 90% and 95%. The important point is that “the program was not introducing new bias but merely reflecting that already in the system” (Lowry and Macpherson, 1988, pg. 657). Additionally, compared to other medical schools in London, this one had more non-European students admitted. Therefore, the “gender-biased” algorithm was less gender-biased than humans at other institutions. Finally, if not for the intervention of this software package, there would have been no evidence linking any prejudices on the part of those who received applications. This is just one such case almost four decades ago.

Para 1 – Current conversation (latest trend/literature)

In times like today, all of us are well familiar with digital computers’ copious information processing capabilities. However, other than that, their facility in manipulating symbols enables them a lot. For example, computer programs can be used for theorem proving in mathematics, converting speech into symbolic form and doing this as well as or better than humans who developed them, playing games at a level better than the creators did, translating texts from one language into another language, composing music and also able to learn from its own mistakes. According to Csaszar & Steinberger (2021) and Shrestha, Ben-Menahem & Von Krogh (2019), artificial intelligence (AI) is currently perceived as a radical technology that has enabled humans to depend on and possibly become overly reliant on intelligences other than human ones. The number of conventional ways that humans have used artificial intelligence (AI) in a business's activity system during the dot-com era and beyond has increased dramatically (Lindebaum, Vesa & Den Hond, 2020; Raisch & Krakowski, 2020). Transportation cars, for example, are one rigid example of how AI is being used in homes through virtual assistance devices, in medicine through artificial intelligence-analyzed scans, in knowledge-related professions through natural language understanding and predictive analysis with the help of generative AIs such as Chat GPT, Co-pilot, Gemini and so on. In short, these programs equip machines to exhibit those behaviours we usually associate with “intelligence” when seen in humans (Minehart,1966). However, in everyday life, automated decisions could result in lower credit scores, reduced payouts on insurance policies, worsened health appraisals, and rejected job applications for deserving candidates in some groups, putting them at a systematic disadvantage (Silberg and Manyika, 2019).

Para 2 – Tension in the current conversation

Today, thanks to advances in machine learning, especially deep learning, the use of artificial intelligence in hiring, performance appraisal, reward and recognition and other managerial functions (Brynjolfsson & McAfee, 2017) has increased and this use of AI is expected to further increase across a wide spectrum of applications and industries. In the next few decades, artificial intelligence will become more pervasive affecting billions of individuals in their various capacities as consumers, employees or members of a community irrespective of geographic coordinates. This means that as we move into the

Commented [NG1]: Statistics of AI:

Commented [NG2R1]: to add

future it will be even more important to understand how AI intersects with bias, and why it does so is this complex. Thus, despite issues in identifying what constitutes unfairness when discussing algorithms, **recently many researchers** (e.g. Raisch & Krakowski, 2021; Turel & Kalhan, 2023; Vanneste & Puranam, 2024) tend to view them with a different lens and as a fresh way forward from long-standing prejudices. Indeed, **recent research** (see Manyika et al., 2019; Orphanou et al., 2022; Van Giffen et al., 2024) agrees that algorithms can mitigate human biases which are inherent in our societies. At the same time, there are fears that they may entrench them deep within society. The available evidence seems to support both arguments demanding serious consideration. Still, these assertions suggest two possibilities: one; using artificial intelligence (AI) models to identify and minimize human biases; two; improving AI systems themselves including data utilization and their development-deployment-use vis-à-vis any tendencies towards perpetuating human and social biases or creating bias themselves or related challenges from such an approach. Thus, there is a debate about whether AI will be less biased than humans in decision-making, as human biases can also influence AI.

Para 3 – How are we addressing the debate?

Commented [AC3]: Can we hint about the methodology here as I have already mentioend about a framework, but what and how of the framework we need to derive from the methodology, I guess.

The aim of our paper is to develop a framework that will help in detecting bias problems in computer systems and address the same. But before exploring more about this issue, it is very important to respond to two potential concerns the reader may have about the framework. First, computer systems can assist in the implementation of social policies that reasonable people may disagree are fair or unjust. Some examples include the employment practices with affirmative action, certain tax laws, and some aspects of federally financed programs. Our framework raises the question of whether computer systems used to implement discriminatory policies embody bias. The answer follows the original (controversial) question: "Is the regulation under focus fair or unfair?" Is the concept of affirmative action, for instance, effective in redressing historical injustices? This article cannot address all of these questions. However, we do assert that if the system's systematic discrimination is found to be unfair, then the charge of bias is justified. Second, bias in computer systems is not inherent in the system itself, but rather a result of its use. An example of emergent bias demonstrates this distinction clearly. Consider an intelligent tutoring system for women's safety, designed for college students. We can assume a high level of literacy without bias. In contrast, designing a system to ensure women's safety in public spaces, such as shopping malls or metro stations, cannot rely on assuming the same literacy level. Less educated individuals may face a disadvantage when using public spaces. Consider the scenario of a technical bias that favours users with shorter-duration jobs over those with longer-duration tasks. Technical bias occurs when participants with short and long-duration employment exceed the system's capacity, rather than being inherent in the program. Practically, bias can be addressed through two types of activities. To identify or "diagnose" bias in any system, we must first identify it. To address bias in systems, we must develop methods to prevent it and correct it when it occurs. We provide some initial directions for this work.

Para 4– The research question and purpose of this paper.

The purpose of the paper is to explore the multifaceted roles of AI in shaping organizational and societal dynamics, the interplay between AI and human decision-making, and the impact of data quality and ethical knowledge on AI's effectiveness and equity.

Commented [AC4]: Can we make it a bit narrow base don the above write up.

We ask the following three research questions:

RQ1: How does AI support or hinder the ability of corporations, international organizations, and social movements to address grand challenges and other social problems (gender bias, cognitive bias, etc.)? How do organizations navigate the bias present in the AI technologies?

RQ2: How does the presence of AI affect the biases and heuristics observed in human decision-making processes within organizations?

Commented [AC5]: We need to combine it into one or modify?

RQ3: To what extent does the generation of new data solve AI's traditional challenges of data availability and quality in decision making?

RQ4: How does active and ethical interventions of knowledge (For instance, Indian spiritual knowledge) in data help mitigate bias present in any community or organisation?

Commented [AC6]: These are good to go.

Despite the promise and potential of AI to improve business operations and performance, generative AI tools still have limitations, such as algorithmic bias. For example, a study published by the U.S. Department of Commerce found that facial recognition AI misidentifies people of color more often than white people.

Supportive nature of AI

AI can analyze large datasets to uncover patterns and insights that may not be visible through traditional methods. This can help organizations identify issues such as systemic gender bias or disparities in access to resources. AI models can forecast trends and potential impacts, allowing organizations to proactively address emerging challenges and tailor their strategies. AI can automate repetitive tasks, freeing up human resources to focus on more strategic and impactful activities. This can improve the efficiency of initiatives aimed at addressing social problems. AI can help in optimizing the allocation of resources by identifying areas where interventions can be most effective. AI can personalize outreach and interventions to better meet the needs of specific groups. For instance, personalized educational tools can address learning disparities or provide targeted support for marginalized communities.

Cross-Disciplinary Insights: AI can integrate and analyze information from various fields, leading to a more comprehensive understanding of complex social problems and informing multi-faceted solutions.

Hinder:

Bias in Training Data: AI systems are often trained on historical data that may contain biases. As a result, these biases can be perpetuated or even amplified in AI outputs, leading to unfair or discriminatory outcomes.

Lack of Contextual Understanding: AI systems may not fully grasp the nuances of social issues, leading to solutions that fail to address underlying causes or exacerbate existing inequalities.

Transparency and Accountability Issues: Many AI models operate as "black boxes," meaning their decision-making processes are not transparent. This lack of clarity can make it difficult to understand how decisions are made and to hold systems accountable for biased outcomes.

Difficulty in Auditing: Assessing and auditing AI systems for bias can be challenging, particularly when proprietary algorithms are involved.

Surveillance and Privacy Concerns: AI technologies can be used in ways that infringe on privacy and civil liberties. For example, AI-powered surveillance can disproportionately target certain groups, exacerbating issues of discrimination and social control.

Navigating AI Bias:

Diverse Data Collection:

Inclusive Datasets: Organizations are increasingly focusing on creating and using diverse and representative datasets to train AI systems. This helps in mitigating bias and ensuring that AI models are more equitable and accurate.

Bias Detection and Mitigation:

Algorithmic Audits: Regular audits and evaluations of AI systems are conducted to identify and address biases. Techniques such as fairness constraints and de-biasing algorithms are employed to improve outcomes.

Transparency Initiatives: Efforts are being made to make AI systems more transparent and interpretable. This includes developing methods for explaining AI decisions and increasing stakeholder involvement in the design and deployment of AI systems.

Ethical AI Frameworks:

Guidelines and Standards: Many organizations and international bodies are developing ethical guidelines and standards for AI development and deployment. These frameworks emphasize fairness, accountability, and transparency.

Cross-Disciplinary Collaboration:

Inclusive Teams: Building diverse teams that include ethicists, social scientists, and community representatives can help in addressing biases and ensuring that AI solutions are more equitable and socially responsible.

Motivation for RQ1, RQ2:

Bias in pre-trained language models : The community has developed a gamut of datasets and methods to measure and mitigate biases in text-only LLMs (Bordia and Bowman, 2019; Liang et al., 2020; Ravfogel et al., 2020; Webster et al., 2020; Lauscher et al., 2021; Smith et al., 2022; Kumar et al., 2023; Nadeem et al., 2021; Nangia et al., 2020). Bias in pre-trained vision models: The use of vision models on various tasks has been hindered by bias in vision, as demonstrated by multiple studies Buolamwini and Gebru (2018); DeVries et al. (2019); Wilson et al. (2019); Rhue (2018); Shankar et al. (2017); Steed and Caliskan (2021). Numerous studies have been conducted to measure the extent of biases present in vision models Steed and Caliskan (2021); Shankar et al. (2017); DeVries et al. (2019); Buolamwini and Gebru (2018).

Bias in Vision and Language models Image-to-text : Hall et al. (2023) introduced a novel portrait based dataset for benchmarking social biases in VLMs for both pronoun resolution and retrieval settings. Srinivasan and Bisk (2021) measure the associations between small set of entities and gender in visual-linguistic models using template based masked language modeling.Zhou et al. (2022); Janghorbani and de Melo (2023) study stereotypes in VLMs.

Text-to-image: Cho et al. (2023) highlights a bias towards generating male figures for job-related prompts and limited skin tone diversity, while probing miniDALL-E Kim et al. (2021) and stable diffusion Rombach et al. (2022b). The prompts used to generate images explicitly specify the profession. Fraser et al. (2023); Ghosh and Caliskan (2023) further highlights stereotypical depictions of people within text-to-image models

On the supportive side, Artifical intelligence is processing a large amount of data to uncover patterns of bias that might be invisible to humans. This can help organizations develop targeted interventions to promote gender equality or reduce cognitive biases. For instance, AI-driven analytics can highlight important patterns, anomalies/ disparities in hiring practices, pay gaps, etc. This will enable more informed decisions to address these issues.

At the same time, AI-based models relies on data and algorithms which may be biased or are not designed with fairness in mind. AI systems trained on biased data can perpetuate or even exacerbate existing inequalities (gender, caste, society, etc.) Moreover, if not carefully managed, AI might prioritize efficiency over ethical considerations and lead to solutions that may address a problem in the short term but ignore deeper societal impacts.

Motivation for RQ3:

Most of the existing methods focus upon generating new data by data augumentation techniques. Generating new data is not a panacea.

Add a brief about data augumentation techniques and what is the process do they follow.

More data does not necessarily mean better data. The relevance and contextual accuracy of the data are equally important. The quality of the new data must be ensured to prevent the introduction of new biases or errors.

AI systems also need to be designed to effectively process and learn from this new data.

Motivation for RQ4:

Integrating ethical and cultural knowledge, such as Indian spiritual principles, into AI data and decision-making processes can help mitigate bias by embedding values of fairness, inclusivity, and respect for diversity into the system.

For example: add the experiments from the traditional Hindu mythology and adding gender-neutrality concept was introduced way before in our traditions.

Add experiments on gender neutral algorithms

Such interventions encourage the design of AI systems that prioritize the well-being of all individuals and communities, rather than focusing solely on efficiency or profit.

For example, Indian spiritual knowledge often emphasizes the interconnectedness of all beings and the importance of ethical conduct.

AI can incorporate these principles, organizations can develop AI systems that are more attuned to social and ethical considerations, thus reducing the likelihood of biased outcomes.

This approach promotes a more holistic understanding of fairness that goes beyond traditional Western concepts of equity.

Related Work

Previous studies explores biases in LLM, OpenAI, GPT2, ChatGPT, LLaama (UNESCO).

Bias in pre-trained language models

The community has developed a gamut of datasets and methods to measure and mitigate biases in text-only LLMs (Bordia and Bowman, 2019; Liang et al., 2020; Ravfogel et al., 2020; Webster et al., 2020; Lauscher et al., 2021; Smith et al., 2022; Kumar et al., 2023; Nadeem et al., 2021; Nangia et al., 2020).

Bias in pre-trained vision models

The use of vision models on various tasks has been hindered by bias in vision, as demonstrated by multiple studies Buolamwini and Gebru (2018); DeVries et al. (2019); Wilson et al. (2019); Rhue (2018); Shankar et al. (2017); Steed and Caliskan (2021). Numerous studies have been conducted to measure the extent of biases present in vision models Steed and Caliskan (2021); Shankar et al. (2017); DeVries et al. (2019); Buolamwini and Gebru (2018).

Bias in Vision and Language models

Image-to-text : Hall et al. (2023) introduced a novel portrait based dataset for benchmarking social biases in VLMs for both pronoun resolution and retrieval settings. Srinivasan and Bisk (2021) measure the associations between small set of entities and gender in visual-linguistic models using template based masked language modeling.Zhou et al. (2022); Janghorbani and de Melo (2023) study stereotypes in VLMs.

Text-to-image: Cho et al. (2023) highlights a bias towards generating male figures for job-related prompts and limited skin tone diversity, while probing miniDALL-E Kim et al. (2021) and stable diffusion Rombach et al. (2022b). The prompts used to generate images explicitly specify the profession. Fraser et al. (2023); Ghosh and Caliskan (2023) further highlights stereotypical depictions of people within text-to-image models

References

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530-1534.

Lowry, S., & Macpherson, G. (1988). A blot on the profession. *British Medical Journal (Clinical research ed.)*, *296*(6623), 657.

Manyika, J., Silberg, J., & Presten, B. (2019). What do we do about the biases in Al. *Harvard Business Review*.

Meinhart, W. A. (1966). Artificial intelligence, computer simulation of human cognitive and social processes, and management thought. *Academy of Management Journal*, *9*(4), 294-307.

Orphanou, K., Otterbacher, J., Kleanthous, S., Batsuren, K., Giunchiglia, F., Bogina, V., ... & Kuflik, T. (2022). Mitigating bias in algorithmic systems—a fish-eye view. *ACM Computing Surveys*, *55*(5), 1-37.

Silberg, J., & Manyika, J. (2019). Notes from the AI frontier: Tackling bias in AI (and in humans). *McKinsey Global Institute*, *1*(6), 1-31.

Turel, O., & Kalhan, S. (2023). Prejudiced against the Machine? Implicit Associations and the Transience of Algorithm Aversion. *MIS Quarterly*, *47*(4).

Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*, 93-106.

Vanneste, B. S., & Puranam, P. (2024). Artificial Intelligence, Trust, and Perceptions of Agency. *Academy of Management Review*, (ja), amr-2022.