Center for Management Science Research
ISSN: 3006-5291 3006-5283 Volume 3, Issue 3, 2025

EXPOSING ISLAMOPHOBIA IN MACHINE LEARNING: A CRITICAL ANALYSIS OF THE EXISTING THEORIES AND BIASES

Dr. Bakht Munir[*1], Dr. Abida Yasin[2], Shahzad Khalid[3], Ume Noreen[4], Syed Baqar Raza Naqvi[5]

[*1]*Postdoctoral Research Fellow, the University of Kansas School of Law, USA*
[2]*Visiting faculty (Law) Rashid Latif Khan University, Lahore*
[3]*PhD Scholar, Brunel University London*
[4]*LLM (University of the Punjab) Assistant Director, Lahore Development Authority (LDA)*
[5]*LLM Scholar University of Lahore (UOL)*

[*1]bakht.munir@ku.edu, [2]abida_yasin5@yahoo.com, [3]shahzad.khalid@brunel.ac.uk,
[4]umenoreenamir@gmail.com, [5]baqarraza83@gmail.com

DOI: https://doi.org/10.5281/zenodo.15173249

Keywords
AI, Religious bias, Islamophobia, Algorithmic bias, disparate impact

Article History
Received on 01 March 2025
Accepted on 01 April 2025
Published on 8 April 2025

Corresponding Author: *

Abstract
With the proliferation of Artificial Intelligence (AI) across different sectors, ethical challenges, such as bias in machine learning, have raised concerns for the AI models' performance. Based on the datasets on which AI systems are trained, AI systems exacerbate and mimic prejudices during the training, where the datasets are inherently biased. Hence, resulting in unfair treatment of individuals based on their gender, race, and religious affiliations. In the digital age, the surge of Islamophobia is a global challenge. Islamophobic bias is a negative feeling towards Muslims, stemming from misinformation about Islam and its followers. The researchers have contributed various mitigating strategies and theories to overcome this challenge. This paper critically investigates the following theories evolved in machine learning and their inherent limitations: (1) Algorithmic Bias, (2) Fairness through Unawareness, (3) Disparate Impact, (4) Equalized Odds and Equal Opportunity, (5) Counterfactual Fairness, and (6) Intersectional Fairness. The research contributes further to mitigating strategies to address the Islamophobic bias in the age of generative AI.

INTRODUCTION

AI refers to the capabilities of machine learning to execute tasks traditionally associated with human intelligence (Allen, 2020; Hernández-Orallo, 2017). AI is rapidly integrating into various fields, transforming every aspect of human life (Boppana, 2022; Kuhar et al., 2024; Rashid & Kausik, 2024). Despite the initial use of AI in the 1950s, there is no consensus on a comprehensive definition of AI due to its versatile functions. In common parlance, AI may denote the machine's ability to make decisions through an evaluative process (Calo, 2017; de Almeida et al., 2021; Turner, 2018). Nevertheless, AI is in its infancy and confronts ethical challenges like bias (Baeza-Yates, 2022; Farmer et al.; Farmer et al., 2024; Oyeniran et al., 2022; Templin et al., 2024; Terra et al., 2023).

Religious biases are unfair treatment and prejudices against individuals or communities based on their religious affiliations. It may exist in several forms, such as discrimination, harassment, stereotype, and intentional exclusion of individuals based on their religious beliefs (Gonsiorek et al., 2009; Roggemans et al., 2015; Ryan & Gardner, 2021; Willard & Norenzayan, 2013). Hence, Islamophobic is a hostile

attitude towards Islam with a belief that Muslims are violent, following their unfair treatment. It is a negative emotion towards Muslims. After the 9/11 incident, the intensification of Islamophobia in the Western world triggered discriminatory surveillance of Muslims, substantiating state actions against Muslim communities, stemming from fabricated information about Islam and their followers (Ayushmaan, 2023; Jaber, 2022; Pratt & Woodlock, 2016; Rifat et al., 2024; Sway, 2005). Considering the context of AI, Islamophobic bias happens when the AI system produces prejudiced outcomes due to a biased dataset or when the AI model is influenced (Aldreabi, 2024; Alkhouri, 2024; Rifat et al., 2024; Samuel-Azran et al., 2024).

Various studies have established that AI tools, such as ChatGPT, often associate Muslims with violence more frequently in comparison to other religions. Since Islam is the second largest religion after Christianity, Islamophobic bias can lead to a global imbalance and could be a potential hazard to embracing AI globally.

1. Relevant Literature

The existing literature containing religious biases is a considerable threat to the widespread proliferation of AI technologies. The more people are exposed to technology, characterized by attributes of automation, the more they decline to hold religious beliefs (Cockrell, March 26, 2024). The Large Language Models (LLMs) used in AI amplify harmful stereotypes, linking Muslims to violence compared to followers of other religions, highlighting the need for new methods to reduce the Islamophobic bias in these models (Abid et al., 2021a). Though the use of adjective techniques helped diminish religious biases, Islamophobic biases are still higher in comparison to other religions (Abid et al., 2021b).

Another study reported that the AI systems generate content like terrorism and bomb blasts in response to words like Islam, Muslims, and mosques. The study further established that while using words associated with other religions, Islamophobic biases were still produced, though the prompts did not contain words about Islam or its followers. The researcher suggested an algorithmic audit for reliability to improve the overall performance of the AI models (Muralidhar, 2021).

A study conducted on the generative capabilities of ChatGPT-3 concluded that GPT unreasonably links Muslims with violence as compared to the other religions based on the principle of "garbage in, garbage out", which means that the training data on the internet will be inspired by whatever biases are presented in the datasets (Samuel, 2021). While highlighting the significance of AI, another study argued that AI could be a potential threat if Muslims show reluctance in adopting AI and could be employed to impart terrorism, radicalization, and apostasy (Khoirunnisa et al., 2023).

Research highlighted that the existing literature has not substantially contributed to mitigating Islamophobic content, leading to their unfair treatment and compromising their safety on social media (Rifat et al., 2024). Another study reported that an increasing number of Islamophobic biases occurred after the 9/11 strategy, alleging Muslims of potentially attacking the sporting events (Samuel-Azran et al., 2024).

Another study necessitated a comprehensive definition of Islamophobia to evade misclassification and to safeguard free speech. The study highlighted the complexity of distinguishing Islamophobia from valid criticism for automated hate detection systems dealing with emotional tones and abusive language. By providing a precise definition of Islamophobia and the deployment of deep learning models, Islamophobic biases can be reduced without compromising valid criticism (Aldreabi & Blackburn, 2023).

2. Limitations in Addressing Islamophobic AI Through Machine Learning

The following are some intrinsic limitations in addressing Islamophobic bias through the process of machine learning: (1) Absence of a Comprehensive definition: The concept of Islamophobia changes across different cultures, making it very challenging to define Islamophobia comprehensively. It necessitates the advancement of AI systems that are both responsive and adaptable to the intricacies of Islamophobic bias. (2) Various kinds and expressions of Islamophobia: Another hindrance in the machine learning process is the various manifestations and expressions of Islamophobia on social media and

online forums, creating a hindrance in the models' training.

For instance, a surge of Islamophobia through online hate speech was reported during the COVID-19 pandemic. Research was conducted on 15656 posts on Facebook across various groups that reported a unique aspect of Islamophobia (Ghasiya & Sasahara, 2022). Likewise, another study on more than seven thousand members across various Facebook groups revealed that certain groups are perpetuating stereotypes against Muslims and reflecting Islam as a violent adversary, resulting in an increasing Islamophobia in the United States.

The following techniques help mitigate Islamophobic bias in machine learning: (1) Universal Language Model Fine-Tuning (ULMFiT): ULMFiT was employed to detect Islamophobic content on Twitter. (Belal et al., 2022). (2) Recurrent Neural Networks (RNNs): By employing Gated Recurrent Units (GRU), RNNs can help detect hate speech (Albadi et al., 2018). (3) Hybrid Model: The hybrid BERT base and CNN model works efficiently in mitigating algorithmic Islamophobic bias (Aldreabi, 2024).

However, a balance between Islamophobia, valid criticism, and the right to speech still needs further appreciation. More importantly, the religious prejudice embodied in the previous datasets through policymaking and adjudication will be replicated by the generative AI because such datasets are apparently neutral but implicitly biased. Religious bias could emerge either to prefer a particular group or community at the cost of others.

3. Modern Conceptions and Theories

The following segment critically examines fundamental concepts and theories related to bias in machine learning, leading to Islamophobic biases and their inherent limitations to address the phenomenon:

3.1. Algorithmic Bias

Algorithmic bias refers to the phenomenon wherein the byproduct of an algorithm unjustifiably favors or disfavors certain groups or individuals in comparison to others, which could lead to unfair decisions, leaving an adverse impact on society. It is a social concept encompassing unfairness, prejudice, and social injustice (Williams et al., 2018). Bias in AI results when two datasets are treated unequally, more likely due to biased hypotheses or inherent prejudices in the training data. Algorithmic bias is exhibited when the generated content is prejudiced based on inaccurate assumptions in the machine learning process due to erroneous algorithmic design, biased data training, or biased input by the developer. The AI system uses real-world data to train itself and replicates the biases accordingly (Sutaria, 2022). To address the issue of algorithmic bias, computational scientists have developed certain techniques (Akter et al., 2021).

Algorithmic bias theory has several limitations that make addressing Islamophobic biases a complex challenge: AI systems learn from historical data. If the training data is biased, the AI models will likely perpetuate those biases. Though the data is unbiased, the algorithms may still favor certain groups at the cost of others based on their structure and decision-making rules. If biased results are fed back into the system, they could amplify existing biases instead of correcting them, reinforcing biases over time. By design, AI systems operate as "black boxes," making it challenging to understand their decision-making processes. Lack of transparency is one of the inherent limitations in recognizing and addressing Islamophobic biases. Strategies like diverse and representative training data and algorithmic audits can help overcome this issue.

3.2. Fairness through Unawareness or Blindness

It refers to an algorithm designed to bypass sensitive features like religion, believing that such attributes' avoidance will result in fair outcomes. It implies the constraints of not expressly using protected characteristics in the decision-making. For instance, fairness may be ensured by not considering gender as a criterion where the decision-making method does not specifically provide for gender (Castelnovo et al., 2022). An algorithm is considered fair unless it uses sensitive features like religion, race, or gender in decision-making, but it can still exhibit bias from other associated attributes. However, in the era of big data, where large databases share common attributes, the theory of fairness through unawareness is not desirable. The active use of sensitive attributes, for

effective surveillance and to avoid discrimination, leads to fair outcomes.

The conventional approach to address bias is known as anti-classification, which strives to achieve fairness through unawareness by omitting sensitive attributes as features from the data. Nevertheless, non-sensitive attributes may establish a relationship with sensitive attributes. For instance, a non-sensitive attribute such as zip code might establish an association with sensitive attributes such as religion and race when many people from the same religious or ethnic background reside in the same vicinity. Backed by high-dimensional and strongly connected datasets, AI systems cannot be fully anticipated in the first place. Despite excluding sensitive attributes, AI may still provide unforeseen links to preserved information based on the complex relationship in the datasets (Ruf & Detyniecki, 2020).

This theory is subject to the following limitations: The AI systems may encounter Islamophobic biases even if religion is excluded, other variables such as name, location, or traditions may act as proxy variables, reintroducing religious biases into the AI models. Similarly, neglecting sensitive features does not address historical biases; for example, Islamophobic biases in societal structures may continue to affect outcomes. Without counting for sensitive features, it becomes difficult to detect, measure, and identify whether the algorithm is producing biases. Disregarding sensitive attributes may lead to decision-making inadvertently disadvantaging Muslims, for instance, a policy aimed to be neutral may still disproportionately affect Muslim communities. To overcome this issue, the algorithm should go beyond the blindness test and encompass fairness frameworks that can ascertain and address discrimination.

3.3. Disparate Impact

Disparate impact refers to a system disproportionately affecting a protected group, though the system seems impartial. Disparate treatment and disparate impact are two models of discrimination based on one theory: The former implies a phenomenon wherein equally qualified people are disproportionately treated based on religion, race, or gender. The latter entails when an employer utilizes a criterion that leaves a disparate impact on a protected group (Willborn, 1984). It refers to an action adversely affecting a protected community, though the action looks fair. This theory aims to ensure that generative AI does not unreasonably affect protected communities that focus on the results rather than the process of recognizing and alleviating passive discrimination that may result in biased outcomes. It is commonly utilized in the hiring process to avoid favoritism and unnecessary bias in the selection of employees. Inadvertent bias is encrypted through disparate impact, which confronts where a selection process perpetuates uneven outcomes for different groups even though it appears unprejudiced. However, there is a shared view regarding the employment of this theory (Calderon, 2024; Feldman et al., 2015; Kassir et al., 2023; Pandey & Caliskan, 2021).

Disparate Impact theory has the following limitations: It does not mandate proof of intent, hence, Islamophobic biases may persist without direct accountability. Biases may be entrenched as cultural norms, making it harder to establish that a particular policy unduly impacts Muslim individuals. Claims under disparate impact often require cogent evidence, which may be challenging to prove, specifically in cases where Islamophobic biases are not explicit. Further, policies can be defended based on security concerns or business necessities. Additional frameworks such as policy revisions, advanced bias detection tools, and cultural sensitivity training are highly advisable to overcome the limitations of disparate impact theory.

3.4. Equalized Odds and Equal Opportunity

Equalized odds, a commonly used measure in supervised learning, imply that protected and non-protected groups should have a common rate for true and false positives (Mehrabi et al., 2021). It is a fairness criterion ascertained through the equilibrium of misclassification rates, false negatives, and false positives across protected groups (Zhang & Bareinboim, 2018). For instance, in a racial or religious context, a Muslim American defendant who would not perpetrate a potential crime will have a parallel opportunity of being discharged in contrast to a non-habitual defendant. Equal opportunity underlines the balance of the true positive rate. It is an ethical method for AI deployment in decision-

making (Chan, 2024). To ensure fairness in machine learning algorithms, the theory of equalized odds and equal opportunity may be employed to overcome discrimination in generative AI (Hardt et al., 2016).

The following are the main limitations of equalized odds and equal opportunity: Though focusing on equalizing error rates, these methods may not essentially address biases, and the AI systems may still perpetuate biases. Even if religion is not explicitly used, other variables may act as proxies, resulting in indirect prejudice. As fairness metrics operate on societal equality and do not account for historical discrimination against Muslim communities, they fail to tackle entrenched societal biases. The employment of this theory may unnecessarily reduce overall accuracy, resulting in fairness adjustments adversely impacting the quality of decision-making. AI systems require more inclusive approaches that can go beyond statistical fairness and actively mitigate religious biases.

3.5. Counterfactual Fairness

Machine learning in predictive decisions can adversely impact certain groups associated with a particular religion or race where the training data embraces inherent biases. A machine learning predictor must consider a religiously biased dataset to overcome biased results. Counterfactual fairness refers to a phenomenon where AI prediction is considered fair if it continues similarly in a counterfactual world where sensitive attributes such as religion or race are distinct. A prediction is fair if it is constant in the real world and in a counterfactual world where the same individual belongs to another gender, ethnic group, or religion. It entails the formation of a hypothetical phenomenon to ascertain whether the AI prediction would remain unchanged based on these attributes (Kusner et al., 2017).

This notion originated from Pearl's causal model, which considers a predictor fair for an individual belonging to a particular group when AI prediction in the real world remains like the counterfactual world where the same individual belongs to another demographic group. However, it is subject to an intrinsic constraint that cannot be evenly quantified from the observational data in similar circumstances owing to the uncertainty of the counterfactual quantity. This limitation can be resolved through an algorithm of the counterfactually fair classifier, which is proven effective for hypothetical and real-world datasets (Wu et al., 2019).

Counterfactual Fairness theory has the following limitations: Fairness assessment may be limited if demographic information is limited, as it requires demographic data to compare individuals across various hypothetical scenarios. It requires causal modeling, making constructing and validating real-world applications hard. Even if various demographic features are disregarded, other correlated attributes in the form of proxies may still perpetuate bias. Counterfactual fairness may compromise predictive accuracy, resulting in undesirable consequences. It can work best in combination with other fairness frameworks for mitigating biases.

3.6. Intersectional Fairness

This theory investigates how coinciding social attributes such as gender, race, and religion could cause discrimination with the ultimate objective of ascertaining fairness across several overlapping attributes. It demonstrates that merit and risk are often motivated by unfair societal processes. It is desired to address this unfairness so individuals may achieve their deserved potential (Foulds et al., 2020). Fairness in AI is inspired by intersectionality – a concept that evolved from the feminist movement that underlines the significance of collectively addressing civil rights and feminism rather than separately (Islam et al., 2023).

An AI system is regarded as fair when the probability of its output is nearly equal across groups defined by various combinations of protected attributes. (Kong, 2022).

In the context of AI, it implies the perception of ensuring fairness in AI algorithms by considering various protected features such as religion, race, and gender intersect and overlap within individuals. In contrast to a conventional approach to fairness that only deems one protected feature, intersectional fairness considers how various attributes can synchronize simultaneously to produce an exceptional experience of discrimination. It aims to ascertain and address biases that may not be evident

through the lens of a single attribute. Hence, fairness criteria should consider various features concurrently and account for their effects (Islam et al., 2023).

Intersectional Fairness theory faces the following limitations: AI systems often struggle with small communities, making it hard to ensure fairness across intersectional identities. It encounters complexity in measurement as traditional fairness metrics stress single attributes, whereas it needs to evaluate multiple dimensions concurrently. Some biases may emerge when the attributes intersect, though they may not exist at the individual attribute level, making the mitigation process more challenging. It may cause the poor performance of the overall predictive accuracy, resulting in undesirable consequences. It demands widespread data collection and algorithmic adjustments, which may not be viable under every situation.

4. Mitigating Strategies

Given the established theories and their inherent limitations, the following are some mitigating strategies that can help diminish the rising issue of Islamophobic biases. As the phenomenon is not the result of a single eventuality, it requires a multi-faceted approach for redressal. (1) Diverse Data: Employing diverse and representative data during data training can help reduce the chances of Islamophobic biases. The data should represent diverse communities, including Muslims. This technique can help prevent biases entrenched in AI systems.

(2) Bias Detection Tools: Employing bias detection tools, such as the AI Fairness 360 toolkit of IBM, during the training phase can significantly contribute to identifying and mitigating religious biases. (3) Ethical Frameworks: If designed to respect religious and cultural diversity, establishing principles and guidelines for AI systems can help reduce religious biases. (4) Inclusive Teams: A multidisciplinary team working on AI systems can help mitigate the chances of biases by fostering a culture of ethics and responsibility.

(5) Community Engagement: Collaborating with Muslim communities to understand their concerns and incorporating their feedback can make the AI systems more responsible and respectful towards their religious sentiments. (6) Accountability and Transparency: Transparency and accountability are mitigating tools to deal with religious bias. Making the AI systems transparent and holding the developers accountable for Islamophobic biases in their AI development helps address biases.

(7) Human-in-the-Loop (HITL): HITL can play a critical role in mitigating Islamophobic biases in AI systems. It involves human oversight at the various levels of the AI lifecycle, ensuring fairness, inclusivity, and accountability. Through human oversight, curated training data can help identify and address Islamophobic content. Human reviewers can monitor the performance of AI models, both at training and production levels, by ensuring that the AI systems are not perpetuating stereotypes or religious biases. Human intervention can help validate or override AI decisions if matters involve religious or cultural inferences. Regular audits of AI systems' performance can help identify and address religious biases that went unnoticed during the developmental stage. Regularly incorporating feedback from diverse users, specifically from the Muslim community, ensures fairness and inclusivity of the AI systems.

5. Conclusion

To conclude, AI systems are prone to Islamophobic biases where the dataset on which these models are trained embeds inherent biases, making a correlation with the training data and the output data. Religious biases may arise if the AI system results in favoring a particular community or targeting at the cost of others based on their religious affiliations. Prejudices and negative stereotypes regarding a community based on their affiliation with Islam replicate Islamophobic biases in AI systems. It is very challenging to debias historical datasets from all kinds of biases. However, various mitigating strategies can help identify and address Islamophobic biases at all levels of AI lifecycles, from training to production levels. Different theories and conceptions have been developed to address religious biases, addressing stereotypes and Islamophobic content. All the theories have certain limitations, adversely affecting the overall performance of the AI systems.

However, a situation may arise where the dataset looks unbiased but may contain intrinsic biases.

The scientifically established mitigating tools may not identify and address inherent biases, necessitating novel mitigating strategies. These strategies include training of diverse and representative data, bias detection tools, ethical frameworks, inclusive teams, community engagement, transparency and accountability, and human oversight. All these mitigating strategies supplement AI capabilities, crafting AI systems that are fairer, ethical, and aligned with societal sentiments, considerably contributing to addressing Islamophobic biases in machine learning.

REFERENCES

Abid, A., Farooqi, M., & Zou, J. (2021a). Large language models associate Muslims with violence. *Nature Machine Intelligence*, *3*(6), 461-463.

Abid, A., Farooqi, M., & Zou, J. (2021b). Persistent anti-muslim bias in large language models. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society,

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. In (Vol. 60, pp. 102387): Elsevier.

Albadi, N., Kurdi, M., & Mishra, S. (2018). Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),

Aldreabi, E., & Blackburn, J. (2023). Enhancing automated hate speech detection: Addressing islamophobia and freedom of speech in online discussions. Proceedings of the International Conference on Advances in Social Networks Analysis and Mining,

Aldreabi, E. A. (2024). *Ethical Algorithms: Safeguarding Freedom of Speech in the Detection of Islamophobic Content* State University of New York at Binghamton].

Alkhouri, K. I. (2024). The role of artificial intelligence in the study of the psychology of religion. *Religions*, *15*(3), 290.

Allen, G. (2020). Understanding AI technology. *Joint Artificial Intelligence Center (JAIC) The Pentagon United States*, *2*(1), 24-32.

Alvi, M., Zisserman, A., & Nellåker, C. (2018). Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. Proceedings of the European conference on computer vision (ECCV) workshops,

Ayushmaan. (2023). Social Media Algorithms in Fuelling Islamophobia: Case Study of Facebook and Youtube. *Jus Corpus LJ*, *4*, 158.

Baeza-Yates, R. (2022). Ethical challenges in AI. Proceedings of the fifteenth ACM international conference on web search and data mining,

Belal, M., Ullah, G., & Khan, A. A. (2022). Islamophobic tweet detection using transfer learning. 2022 International Conference on Connected Systems & Intelligence (CSI),

Bi, Q., Goodman, K. E., Kaminsky, J., & Lessler, J. (2019). What is machine learning? A primer for the epidemiologist. *American journal of epidemiology*, *188*(12), 2222-2239.

Boppana, V. R. (2022). Machine Learning and AI Learning: Understanding the Revolution. *Journal of Innovative Technologies*, *5*(1).

Calderon, V. (2024). Unintentional Algorithmic Discrimination: How Artificial Intelligence Undermines Disparate Impact Jurisprudence. *Duke Law & Technology Review*, *24*(1), 28-51.

Calo, R. (2017). Artificial intelligence policy: a primer and roadmap. *UCDL Rev.*, *51*, 399.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, *12*(1), 4209.

Chan, G. K. (2024). AI employment decision-making: integrating the equal opportunity merit principle and explainable AI. *AI & SOCIETY*, *39*(3), 1027-1038.

Cockrell, J. (March 26, 2024). Where AI Thrives, Religion May Struggle. *CBR - Artificial Intelligence*. https://www.chicagobooth.edu/review/where-ai-thrives-religion-may-struggle

de Almeida, P. G. R., dos Santos, C. D., & Farias, J. S. (2021). Artificial intelligence regulation: a framework for governance. *Ethics and Information Technology*, *23*(3), 505-525.

Farmer, R. L., Lockwood, A. B., Goforth, A., & Christopher Thomas, J. CHALLENGES, AND ETHICAL CONSIDERATIONS.

Farmer, R. L., Lockwood, A. B., Goforth, A., & Thomas, C. (2024). Artificial intelligence in practice: Opportunities, challenges, and ethical considerations. *Professional Psychology: Research and Practice*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining,

Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020). An intersectional definition of fairness. 2020 IEEE 36th International Conference on Data Engineering (ICDE),

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, *4*(1), 1-58.

Ghasiya, P., & Sasahara, K. (2022). Rapid sharing of Islamophobic hate on Facebook: The case of the Tablighi Jamaat controversy. *Social Media+ Society*, *8*(4), 20563051221129151.

Goel, A., Goel, A. K., & Kumar, A. (2023). The role of artificial neural network and machine learning in utilizing spatial information. *Spatial Information Research*, *31*(3), 275-285.

Gonsiorek, J. C., Richards, P. S., Pargament, K. I., & McMinn, M. R. (2009). Ethical challenges and opportunities at the edge: Incorporating spirituality and religion into psychotherapy. *Professional Psychology: Research and Practice*, *40*(4), 385a.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

Hernández-Orallo, J. (2017). Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, *48*, 397-447.

Islam, R., Keya, K. N., Pan, S., Sarwate, A. D., & Foulds, J. R. (2023). Differential fairness: an intersectional framework for fair AI. *Entropy*, *25*(4), 660.

Jaber, N. (2022). Islamophobia: definition, history, and aspects. *Nazhruna: Jurnal Pendidikan Islam*, *5*(2), 327-338.

Kassir, S., Baker, L., Dolphin, J., & Polli, F. (2023). AI for hiring in context: a perspective on overcoming the unique challenges of employment research to mitigate disparate impact. *AI and Ethics*, *3*(3), 845-868.

Khoirunnisa, A., Rohman, F., Azizah, H. A., Ardianti, D., Maghfiroh, A. L., & Noor, A. M. (2023). Islam in the Midst of AI (Artificial Intelligence) Struggles: Between Opportunities and Threats. *SUHUF*, *35*(1), 26-30.

Kim, B., Kim, H., Kim, K., Kim, S., & Kim, J. (2019). Learning not to learn: Training deep neural networks with biased data. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,

Kong, Y. (2022). Are "intersectionally fair" ai algorithms really fair to women of color? a philosophical analysis. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency,

Kuhar, N., Kumria, P., & Rani, S. (2024). Overview of Applications of Artificial Intelligence (AI) in Diverse Fields. In *Application of Artificial Intelligence in Wastewater Treatment* (pp. 41-83). Springer.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, *30*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM *Computing Surveys (CSUR)*, *54*(6), 1-35.

Muralidhar, D. (2021). Examining religion bias in AI text generators. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society,

Oyeniran, C., Adewusi, A. O., Adeleke, A. G., Akwawa, L. A., & Azubuko, C. F. (2022). Ethical AI: Addressing bias in machine learning models and software applications. *Computer Science & IT Research Journal*, *3*(3), 115-126.

Pandey, A., & Caliskan, A. (2021). Disparate impact of artificial intelligence bias in ridehailing economy's price discrimination algorithms. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society,

Pratt, D., & Woodlock, R. (2016). Introduction: Understanding Islamophobia. *Fear of Muslims? International Perspectives on Islamophobia*, 1-18.

Rashid, A. B., & Kausik, A. K. (2024). AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. *Hybrid Advances*, 100277.

Rifat, M. R., Asha, A. Z., Jadon, S., Yan, X., Guha, S., & Ahmed, S. I. (2024). Combating Islamophobia: Compromise, Community, and Harmony in Mitigating Harmful Online Content. *ACM Transactions on Social Computing*, *7*(1), 1-32.

Roggemans, L., Spruyt, B., Droogenbroeck, F. V., & Keppens, G. (2015). Religion and negative attitudes towards homosexuals: An analysis of urban young people and their attitudes towards homosexuality. *Young*, *23*(3), 254-276.

Ruf, B., & Detyniecki, M. (2020). Active fairness instead of unawareness. *arXiv preprint arXiv:2009.06251*.

Ryan, A. M., & Gardner, D. M. (2021). Religious harassment and bullying in the workplace. *Dignity and inclusion at work*, 463-487.

Samuel-Azran, T., Manor, I., Yitzhak, E., & Galily, Y. (2024). Analyzing AI Bias: The Discourse of Terror and Sport Ahead of Paris 2024 Olympics. *American Behavioral Scientist*, 00027642241261265.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, *3*(3), 210-229.

Samuel, S. (2021). AI's Islamophobia problem. *Vox. September*, *18*, 2021.

Sutaria, N. (2022). Bias and ethical concerns in machine learning. *ISACA J.*, *4*, 1-4.

Sway, M. A. (2005). Islamophobia: Meaning, Manifestations, Causes. *Palestine-Israel Journal of Politics, Economics & Culture*, *12*.

Templin, T., Perez, M. W., Sylvia, S., Leek, J., & Sinnott-Armstrong, N. (2024). Addressing 6 challenges in generative AI for digital health: A scoping review. *PLOS Digital Health*, *3*(5), e0000503.

Terra, M., Baklola, M., Ali, S., & El-Bastawisy, K. (2023). Opportunities, applications, challenges and ethical implications of artificial intelligence in psychiatry: a narrative review. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, *59*(1), 80.

Turner, J. (2018). *Robot rules: Regulating artificial intelligence*. Springer.

Willard, A. K., & Norenzayan, A. (2013). Cognitive biases explain religious belief, paranormal belief, and belief in life's purpose. *Cognition*, *129*(2), 379-391.

Willborn, S. L. (1984). The disparate impact model of discrimination: Theory and limits. *Am. UL Rev.*, *34*, 799.

Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *Journal of Information Policy*, *8*, 78-115.

Wu, Y., Zhang, L., & Wu, X. (2019). Counterfactual fairness: Unidentification, bound and algorithm. Proceedings of the twenty-eighth international joint conference on Artificial Intelligence,

Zhang, J., & Bareinboim, E. (2018). Equality of opportunity in classification: A causal approach. *Advances in neural information processing systems*, *31*.