Course: TIMG 5204A: Responsible AI Ethics [S23]

Assignment #3: Trustworthy AI Assessment

Topic: Facebook Ad's algorithm

By Group #2:

Ahnaf Mohsin

Arezoo Khojasteh Abbasi

Manas Agarwal

Maryam Agha Hosseinalishirazi

Rajdeep Jinegar

Shahrzad Kahrizi

Date of Submission: June 16th, 2023

Table of Content

Section 1: Introduction

In our analysis, we will examine the trustworthiness of Facebook's ad algorithm, focusing on the article titled "On Assessing Trustworthy AI in Healthcare: Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls" published in Frontiers in Human Dynamics [1], along with other relevant articles. Our evaluation will be guided by the principles outlined in the EU Framework for Trustworthy AI and insights from the Z-inspection® process.

The objective of this analysis is to identify key areas where the Facebook ad algorithm can be improved to enhance trustworthiness. The EU Framework for Trustworthy AI provides a comprehensive set of principles for responsible AI development. These principles include transparency, accountability, fairness, and non-discrimination, among others. By aligning with these principles, Facebook can ensure that its ad algorithm operates in a transparent and fair manner, without promoting discriminatory practices.

In addition to the EU Framework, the Z-inspection process emphasizes transparency, accountability, and robustness in AI systems. This process involves auditing AI algorithms and their deployment to ensure ethical practices are followed. By incorporating insights from the Z-inspection process, Facebook can enhance the reliability and trustworthiness of its ad algorithm. The Z-Inspection process highlights the significance of continuous evaluation and improvement throughout the algorithm's life cycle. This approach allows Facebook to mitigate risks, prevent biases, and promote fairness and user privacy.

By combining the principles of the EU Framework for Trustworthy AI and the Z-Inspection process, Facebook can cultivate a safer and more reliable digital advertising platform. Through collaboration with policymakers, platforms, and advertisers, a fair and inclusive digital advertising ecosystem can be established, contributing to the building of trust among users and stakeholders. Thus, by leveraging the concepts discussed in the aforementioned article, we aim to assess Facebook's ad algorithm and shed light on its adherence to these critical principles.

In the following sections, we will delve deeper into the evaluation of Facebook's ad algorithm based on the principles outlined in the EU Framework for Trustworthy AI using the Z-Inspection process and its guidelines for ethical development.

Section 2: Analysis using Z-Inspection

To assess the trustworthiness of Facebook's ad algorithm, we will employ a comprehensive evaluation methodology. Assessing Facebook Ad's Algorithm's "trustworthiness" based on the EU "Framework for Trustworthy AI" using the Z-inspection process requires evaluating its adherence to its three ethics guidelines (lawful, ethical, and robust), four ethical principles (respect for human autonomy, prevention of harm, fairness, and explicability), and its seven requirements. We will then discuss how the process can be applied to each stage of its life cycle (design, development, deployment, monitoring), enabling us to translate its principles into tangible practices [1] [2].

1. Z-Inspection's three ethics guidelines for the trustworthy system:

(i) Lawful: Facebook ad's algorithm should prioritize compliance with all applicable laws and regulations governing advertising and data privacy. This includes adhering to anti-discrimination laws, privacy regulations, and any other relevant legislation in the jurisdictions where the algorithm is deployed. By ensuring the algorithm operates within legal boundaries, Facebook can mitigate the risk of engaging in unethical practices and avoid potential legal repercussions.

(ii) Ethical: The process should uphold ethical principles and values that go beyond legal requirements. Facebook should establish a clear set of ethical guidelines to guide the development, deployment, and use of the advertising algorithm. These guidelines should encompass principles such as fairness, accountability, transparency, and respect for user autonomy. Ethical considerations should be integrated into the algorithm's design, training, and decision-making processes to ensure it aligns with societal expectations and values.

(iii) Robust: The algorithm should be evaluated for technical robustness, including reliability, accuracy, and resistance against adversarial attacks. Rigorous testing and validation procedures are necessary to ensure consistent and reliable performance across various scenarios. Furthermore, evaluating potential biases, discrimination, and unintended consequences is crucial, and the algorithm should be adaptable to evolving societal norms and expectations.

By incorporating these ethics guidelines into the assessment process, Facebook can ensure that its ad's algorithm not only meets legal requirements but also upholds ethical principles and values.

2. Z-Inspection's four ethical principles:

(i) Respect for human autonomy: Facebook Ad's Algorithm should respect the autonomy of users by ensuring that the ads delivered to them are not influenced by factors that violate their autonomy, such as their race, gender, or age. Based on the provided report, it is evident that the algorithm exhibits demographic biases, behavioural biases, proxy biases, and algorithmic fairness biases. These biases can infringe upon the principle of respect for human autonomy, as they may lead to unequal treatment and exclusion of certain individuals or groups based on protected attributes.

(ii) Prevention of harm: The algorithm should be designed and implemented in a way that minimizes the potential harm it may cause to individuals or groups. Biases in Facebook's ad algorithm, such as reinforcement of stereotypes, economic inequities, discriminatory practices, and psychological effects, can lead to harm by perpetuating inequalities, exclusion, and unfair treatment. To prevent harm, Facebook needs to take proactive measures to identify and mitigate biases and establish clear guidelines and policies for fair and inclusive ad delivery.

(iii) Fairness: It is a paramount principle in evaluating the trustworthiness of AI systems, ensuring the avoidance of unjustified bias and promoting equal treatment and opportunities. Within Facebook's ad algorithm, biases such as demographic, interest-based, and feedback loop biases have been identified and must be actively addressed. Algorithmic improvements, refinement of machine learning models, and the incorporation of additional fairness metrics are necessary to mitigate these biases. Furthermore, fairness testing tools can aid advertisers in identifying and rectifying unintended biases in their ad campaigns, fostering fairness in targeted advertising.

(iv) Explicability: The algorithm must be transparent and explainable to users and stakeholders, allowing them to understand how their data is used for ad targeting and to detect any potential biases or unfair practices. However, the proprietary nature of Facebook's algorithm makes it difficult to evaluate its level of explicability. To address this, Facebook should offer clearer guidelines, share comprehensive information about ad delivery factors, and consider external audits to improve explicability.

Thus, Facebook needs to actively address and mitigate biases in its algorithm to prioritize user autonomy, prevent harm, promote fairness, and provide explicability to users and stakeholders.

3. Z-Inspection's seven requirements for trustworthy AI which need to be continuously evaluated and addressed through the AI system's life cycle:

i) Human Agency and Oversight: The Z-Inspection process should ensure that humans maintain control and oversight over the advertising algorithm. While the algorithm may automate certain decision-making processes, there should be appropriate human intervention and final decision-making authority. Humans should be able to understand, challenge, and override algorithmic decisions when needed.

(ii) Technical robustness and safety: This involve assessing the algorithm's reliability, performance, and resilience to failures or attacks. Measures should be implemented to prevent unintended behaviors, mitigate risks, and ensure the algorithm operates within acceptable bounds of safety and performance. Reliability is a fundamental consideration in evaluating the trustworthiness of AI systems. While the specific details of Facebook's algorithm are proprietary, continuous evaluation and improvement are vital to maintaining reliability. Ongoing research and development efforts should be undertaken to enhance the fairness of the algorithm, refine machine learning models, and reduce discriminatory outcomes. Additional articles, such as "Ensuring Reliability in AI Systems: Challenges and Best Practices," [3] can provide in-depth analysis and recommendations for ensuring the technical robustness and safety of AI systems.

iii) Privacy and Data Governance: Facebook should prioritize privacy and data governance principles, and ensure that user data used by the advertising algorithm is handled in accordance with relevant privacy laws and regulations. Data collection, storage, and processing practices should prioritize user consent, data minimization, and secure handling to protect individual privacy rights.

(iv) Transparency: It plays a pivotal role in evaluating the trustworthiness of AI systems. It involves providing clear information and explanations about the behaviour and outcomes of the system. In the context of Facebook's ad algorithm, transparency can be improved by offering more comprehensive guidelines for advertisers and users regarding ad targeting. Additionally, the algorithm's inner workings should be made accessible for external scrutiny. While Facebook has taken some steps in this direction by sharing insights into factors influencing ad delivery, further efforts are required to enhance transparency. Insights from related articles, such as "Ensuring

Transparency and Explainability in AI Systems," [4] can be leveraged to inform the analysis and provide a more comprehensive understanding of transparency requirements.

(v) Diversity, non-discrimination, and fairness: These are critical principles in trustworthy AI, ensuring that the AI system does not unfairly disadvantage or exclude individuals or groups based on protected attributes. Facebook must prioritize ethical considerations throughout the ad delivery process. This includes adopting algorithmic fairness techniques and incorporating fairness and inclusivity as core principles in the design and implementation of ad delivery algorithms. Promoting a culture of fairness within the organization is vital to prevent discriminatory practices. Articles like "Mitigating Bias in Algorithmic Hiring: A Systematic Review" [5] can provide insights into the challenges and strategies for addressing discrimination in AI systems.

vi) Societal and Environmental Wellbeing: Facebook should evaluate the algorithm's potential social, economic, and environmental consequences, taking into account factors such as user well-being, community cohesion, and environmental sustainability. Mitigation measures should be implemented to minimize negative externalities and promote positive social outcomes.

(vii) Accountability: It is another crucial aspect of trustworthy AI. It entails defining responsibilities and mechanisms for the development, deployment, and outcomes of the AI system. To enhance accountability, Facebook can conduct regular audits and assessments of the algorithm's fairness and bias. Collaborating with external organizations for independent audits would contribute to establishing a higher level of accountability and increasing public trust. Incorporating insights from articles such as "Accountability in Artificial Intelligence: A Legal Perspective" [6] can provide valuable guidance on the legal and ethical aspects of accountability in AI systems.

Thus, by continuously evaluating and addressing these seven requirements for trustworthy AI throughout the advertising algorithm's life cycle, Facebook can enhance user trust, mitigate risks, and ensure that the algorithm aligns with societal expectations and values. Regular monitoring, audits, and stakeholder engagement are essential to assess compliance with these requirements and drive ongoing improvements.

4. Z-Inspection process for the AI's life cycle:

(i) Design: During the design phase, it is important to consider ethical principles and potential biases. Facebook should incorporate diverse perspectives into the algorithm's development process to identify and rectify biases. They should also establish clear guidelines and policies for fair and inclusive ad delivery practices. Ethical considerations, such as respect for human autonomy and prevention of harm, should be prioritized from the beginning.

(ii) Development: In the development phase, Facebook should invest in research and development to enhance the fairness of the algorithm. This may involve refining the machine learning models, incorporating additional fairness metrics, and conducting ongoing analysis to detect and mitigate biases. Algorithmic improvements should align with ethical principles, and fairness testing tools should be introduced to allow advertisers to assess the fairness of their ad campaigns.

(iii) Deployment: When deploying the algorithm, Facebook should prioritize transparency and accountability. They should provide advertisers with more information about how their ads are delivered, establish clear guidelines and policies for fair and inclusive ad delivery practices, and incorporate accessibility guidelines into the algorithm. External audits can also be conducted to assess potential biases and provide valuable insights and recommendations for improvement

(iv) Monitoring: Continuous monitoring is crucial to ensure that the algorithm is functioning ethically and in accordance with the established principles. Facebook should implement robust monitoring systems to detect and address biases in real time. Regular audits and evaluations, both internal and external, can provide independent assessments of the algorithm's performance and identify areas for improvement. User feedback and input should also be actively sought and considered to address any concerns or issues that arise.

Overall, the Z-inspection process can help assess the trustworthiness of Facebook Ad's Algorithm by evaluating its adherence to ethical principles and its performance at each stage of the AI life cycle. By incorporating diverse perspectives, investing in algorithmic improvements, promoting transparency, and conducting regular audits, Facebook can work towards building a more trustworthy and fair advertising platform.

Section 3: Conclusion

In conclusion, our analysis draws upon the article "On Assessing Trustworthy AI in Healthcare" [1] and other relevant literature to evaluate the trustworthiness of Facebook's ad algorithm based on the principles outlined in the EU Framework for Trustworthy AI. By enhancing transparency, accountability, fairness, reliability, and non-discrimination, Facebook can address the biases within the algorithm, promote a more inclusive digital advertising landscape, and foster trust among users and stakeholders. Ongoing efforts, such as regular audits, algorithmic improvements, and collaboration with external organizations, are essential to continuously evaluate, monitor, and improve the ad algorithm, ensuring the mitigation of biases and promoting fairness. Policymakers, platforms, and advertisers must collaborate to create a fair and inclusive digital advertising ecosystem.

Furthermore, the implementation of the Z-Inspection process by Facebook serves as a valuable mechanism for ensuring the ethical development and deployment of its advertising algorithm. By adhering to the three key principles of transparency, accountability, and robustness, Facebook can promote trust and confidence among its users and stakeholders. The Z-Inspection process incorporates a comprehensive set of guidelines and requirements, including legal compliance, ethical considerations, technical robustness, and societal well-being. By continuously evaluating and addressing these aspects throughout the algorithm's life cycle, Facebook can mitigate risks, prevent biases, and promote fairness and user privacy. Thus, by leveraging the insights from the article on ‘trustworthy AI in healthcare’ [1] and other relevant articles, we extend the principles to the domain of digital advertising, emphasizing the significance of responsible and ethical AI practices in building trust among users and stakeholders.

Section 4: References

[1] R. V. Zicari et al., “On assessing trustworthy AI in healthcare. machine learning as a supportive tool to recognize cardiac arrest in emergency calls,” Frontiers, https://www.frontiersin.org/articles/10.3389/fhumd.2021.673104/full (accessed Jun. 16, 2023).

[2] Dr. S. Nikpoor, Class Lectures, Topic: "Responsible AI Ethics" TIMG 5204, Carleton University, Ottawa, ON, Canada (accessed Jun. 16, 2023)

[3] C. Science, “How do you ensure the quality and reliability of Your Machine Learning and Artificial Intelligence Systems?,” Best Practices for Reliable Machine Learning and AI Systems, https://www.linkedin.com/advice/3/how-do-you-ensure-quality-reliability-your-machine (accessed Jun. 16, 2023).

[4] Nagadivya Balasubramaniam et al., “Transparency and explainability of AI systems: From ethical guidelines to requirements,” Information and Software Technology, https://www.sciencedirect.com/science/article/pii/S0950584923000514 (accessed Jun. 16, 2023).

[5] M. R. C. University et al., “Mitigating bias in algorithmic hiring: Proceedings of the 2020 conference on fairness, accountability, and transparency,” ACM Conferences, https://dl.acm.org/doi/10.1145/3351095.3372828 (accessed Jun. 16, 2023).

[6] F. Doshi-Velez et al., “Accountability of AI under the law: The role of explanation,” arXiv Vanity, https://www.arxiv-vanity.com/papers/1711.01134/ (accessed Jun. 16, 2023).