SIT 799

Task 2.2: Philosophical essay on Ethics in AI

Racist robots. What is the impact of racists robots on our society and how do we eliminate AI racism?

With increased computational capabilities theoretical concepts of AI are now being applied to multiple verticals of the society. Increase in organically generated data acts as a fuel for this advancement, huge amount of organically generated data is now being used to train models which are powering cutting edge innovation. The models trained using data are susceptible to inheriting biases that are present in this data, which leads to AI solutions that appear to be racist. In this essay, I will examine the question "Racist robots. What is the impact of racists robots on our society and how do we eliminate AI racism?

To examine the biases effectively and critically in AI solutions, we need to examine the biases which these solutions inherit from data. This essay first examines how visual data can lead to computer vision solutions to be racist, in the next section this essay explores the biases present in textual data generated from various sources and highlights how several words are linked with biases originating on the base of gender, ethnicity and religion. In this section, this essay also highlights how speech-based systems are prone to inheriting human biases from data. This essay also showcases how such biases can be avoided and ways to develop a near neutral AI system.

Since Viola-Jones [1] cascade-based face detectors, face detection algorithms have come a long way. Recent face recognition systems such as FaceNet [2] and ArcFace [3] now boast of near human level face recognition accuracy with FaceNet-512 having an accuracy of 99.65% on Labelled Faces in the wild (LFW) [4] dataset. In a critical system like face recognition, it is essential that the system does not misidentify individuals which can lead to false incrimination or accusation. Accurate detection of faces coupled with state-of-the-art face recognition technology is now being used by businesses and governments alike for safety and surveillance [5]. The bias in such solutions arise from the fact that there is an uneven representation of different subgroups of society. This leads to the system performing better for a specific subgroup while having a poor performance relatively for minor subgroups that are not represented in the dataset.

A recent incident shows that Detroit police falsely arresting an individual of color [6] when the individual was mismatched because of a fault in the face recognition system. An NIST study [7] where the study on 189 facial recognition algorithms reported that most of the facial recognition algorithms display inherent bias. The study reports that the algorithms falsely identified Black and Asian faces 10 to 100 times more that they did with white faces. The legislative reforms to monitor such technologies also lack a robust framework which leads to unwanted incidents. These incidents are then met with regressive measures such as banning the use of facial recognition technologies altogether, while these measures do handle the situation, they act like a quick fix to a much more complex problem. For sustainable development and continuous technological advancements, it is essential that technological reforms are suggested to tackle these issues, [8] offer multiple solutions to problems of inherent bias in face recognition technologies. These solutions include:

- Improved Data preparation: Using an evenly distributed data to include minorities and marginalised subgroups.
- Public Dataset Auditing: Using public and external agencies to audit the dataset on which the model is being trained to identify biases that may creep into the data.

- Lowering Algorithmic Bias: This is a newer phenomenon where debiasing algorithms are being used to tackle algorithmic biases.

Recent measures from companies like Google include displaying a model card with each such model that includes information of accuracy and other evaluation metrics specific to ethnic subgroups. These measures point to a promising direction which can lead to a broader use and adaptation of face recognition technologies.

While the previous section elaborates on how face recognition and computer vision systems are susceptible to algorithmic biases arising from data, it is essential to highlight that AI systems based on speech data as well as textual data are also prone to such biases which may lead to a racist AI solution. [9] presents an example scenario where state-of-the-art Automatic Speech Recognition solutions are prone to the issues of bias. This is observed when the solutions perform better at identifying certain accents and pronunciations and perform poorly on accents which are not represented well in data. This leads to a sub optimal user experience for groups which are not represented in the dataset, the study also uncovers the bias that such ASR systems do not consider users with speech impairments. Similarly, textual data is also prone to such biases which lead to chatbots and NLP solutions displaying a racist behaviour. We can observe in [10] how dataset scrapped from Reddit (this is true for all organically generated data on the internet) show human biases, which lead to NLP systems mapping negative words based on ethnicity, gender, and religion. The paper also offers an effective way to handle such biases by using debiasing algorithms specific to textual data.

This essay focuses on highlighting biases through case studies of face recognition scenarios and offers preventive measures to avoid such situations. We have also seen that such biases are not only prevalent in visual data but also can be found in other organically generated data types such as audio and textual data. The unifying factor across all data sources is the presence of human factor, AI solutions in themselves are neither racist nor biased, they just learn to model human biases which are present in data. The solutions to these racist AI solutions are also common across all data sources, if corrective infrastructure is placed at data collection and annotation stage then a huge chuck of data bias can be handled. Coupled with rigorous and thorough testing these methods can aid in developing a bias free and neutral AI solutions.

References

[1] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2005.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] J. Deng, J. Guo, J. Yang, N. Xue, I. Cotsia, and S. P. Zafeiriou, “ArcFace: Additive Angular Margin Loss for deep face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, pp. 1–1, 2021.

[4] B. Gary, M. Huang, T. Ramesh, and E. Berg, *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. .

[5] S. Selvi, D. Sivakumar, S. Sowmiya, K. Suba, and Raja, “Face recognition using Haar-cascade classifier for criminal identification,” 2019.

[6] B. Rauenzahn, “Facing bias in facial recognition technology,” *The Regulatory Review*, 20-Mar-2021. [Online]. Available: https://www.theregreview.org/2021/03/20/saturday-seminar-facing-bias-in-facial-recognition-technology/. [Accessed: 10-Aug-2022].

[7] C. Boutin, “NIST study evaluates effects of race, age, sex on face recognition software | NIST,” 2019.

[8] J. Lunter, “Beating the bias in facial recognition technology,” *Biom. technol. today*, vol. 2020, no. 9, pp. 5–7, 2020.

[9] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, “Quantifying bias in automatic speech recognition,” *arXiv [eess.AS]*, 2021.

[10] T. Manzini, Y. C. Lim, Y. Tsvetkov, and A. W. Black, “Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings,” *arXiv [cs.CL]*, 2019.