BIAS IN ARTIFICIAL INTELLIGENCE: AN ANALYSIS WITH SPECIAL FOCUS ON AUTOMATED CONTENT MODERATION

Rithika Reddy Shyamala, BA LLB (Hons) Student at O.P. Jindal Global Law School

ABSTRACT

The extensive use of AI and machine-learning systems in our day-to-day life, has made it imperative for us to discuss the fairness issues in relation to such AI.[1] It is a common assumption that Artificial Intelligence systems or Machine Learning algorithms are fair, just and unbiased in making decisions because they are meant to interject where human biases exist.[2] But this is a mere delusion. AI systems are plagued with the negative effect of bias amongst many other things. Bias is reflected in the actions and decisions of the AI system, against an individual or a particular section of the society. The aim of this paper is to discuss the concept of bias in AI, how it impacts online platforms where AI is heavily being used for content moderation and how such bias can be tackled or mitigated.

[1] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Argam Galystyan, *A survey on Bias and Fairness in Machine Learning*, USC-INFORMATION SCIENCES INSTITUTE (June 15, 2022) https://arxiv.org/pdf/1908.09635.pdf

[2] *How to Reduce Bias in AI*, APPEN.com (June 15, 2022) https://appen.com/blog/how-to-reduce-bias-in-ai/

The internet is being used by an approximate of 4.5 billion users who are creating content in the form of billions of images, video, messages, posts etc. There is a need to regulate this content to ensure that the internet is a safe and positive experience. This regulation takes place through a process called content moderation which is defined as an "*organized practice of screening online content based on the characteristics of the website, its targeted audience, and jurisdictions of user-generated content (UGC), to determine whether the UGC is appropriate*."[3] Initially, content moderation was carried out by humans where the human moderator was required to evaluate every single post and screen whether it complied with the company's guidelines. "*With the growth in the amount of content posted online, as well as the public and regulatory pressure on platforms to protect users and expeditiously remove illicit content, it has become almost impossible for online platforms to rely exclusively on human reviewers*."[4] However, with the growth in the number of posts and content on the internet, human moderators are not being able to solve the issue efficiently and effectively. Thus, artificial intelligence has been proposed as an alternative to human moderators. AI is involved in proactive *detection* of content where it detects potentially problematic posts and automated *evaluation* of that content. After evaluation, AI tools "*enforce the decision about whether a post violates the host's content policy (or the law) by automatically removing or demoting content*."[5][4]

One of the biggest benefits of using AI based content moderation is that content can be reviewed and filtered automatically at a much faster and efficient rate. Since the presence of harmful and toxic content on public platforms, even for a short time period, can be very detrimental to the society, immediate action to remove it becomes critical. This is where AI comes in handy. Inappropriate content is flagged and stopped from being posted or deleted from the platform almost instantaneously which is not possible when the task is to be performed by human reviewers. AI thereby acts as the perfect support mechanism for human moderators by helping them in making the process speedy, accurate and efficient. But this is only the good side to the use of AI, one of the biggest negatives and the main element of discussion in this paper is, contrary to the popular belief, is the Bias in AI.

[3] Yifat Nahmias, Maayan Perel, *The Oversight of Content Moderation by AI: Impact Assessments and their Limitations*, 58 (1) Harv. J. on Legis., 2020 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3565025
[4] ibid.
[5] Emma J Llanso, *No Amount of "AI" in Content Moderation Will Solve Filtering's prior Restraint Problem*, BIG DATA AND SOCIETY, SAGE JOURNALS (June 15, 2022) https://journals.sagepub.com/doi/full/10.1177/2053951720920686

There is ample evidence of Bias in AI.[6] While focusing on the goods of AI – efficiency, accuracy and cost-effectiveness,[7] we tend to ignore the fact that algorithms are actually designed and assembled by humans. So, if humans can be biased, then so can AI. Bias in AI can be of many types. This paper will aim to discuss the two major types of bias, i.e. algorithmic bias and societal or human bias. The former, also referred to as "data bias" is caused majorly due to algorithms being trained using biased data.[8] While the latter is an effect of the assumptions and norms we have as humans or as a society that lead us to having certain blind spots in our thinking.[9] Societal bias also impacts algorithmic bias because AI doesn't exist in a vacuum and the algorithms are devised by the same people who have certain biases and as a result the AI tends to think the way it is taught.[10] This is only logical because if AI can adopt the reasoning capacity of human beings, then it is inevitable that it could make certain mistakes and misjudgments like humans as well.[11]

Bias can enter AI systems in multiple different ways. One of the most prominent ways in which bias creeps into algorithms is through training data. It is very important to ensure that the training data being used to train the AI shall be free from embedded bias so as to build an unbiased algorithm.[12] AI "learns" what data should be allowed on the internet and what ought to be removed by examining vast amounts of data and the datasets they are trained on. The decisions taken by AI are hugely dependent on the nature of the datasets. For example, when AI is removing a post because it contains nudity, it recognises nudity by analysing the "*proportion of pixels in an image that fall into a specific color range that has been pre-identified as representing skin color*."[13] The problem with this tool is that the AI may be vulnerable to misclassification of underrepresented skin tones which the training data did not 'teach' AI about. Thus, if the training data just had images of people with fair tones while

---

[6] Gregory S. Nelson, *Bias in Artificial Intelligence*, 80 (4) NC Med. J., Page 220 (June 15, 2022) https://www.ncmedicaljournal.com/content/ncm/80/4/220.full.pdf

[7] Genevive Smith, Ishita Rustagi, *Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook*, Berkeley Haas Center for Equity, Gender and Leadership, Page 16, 2020 https://haas.berkeley.edu/wpcontent/uploads/UCB_Playbook_R10_V2_spreads2.pdf

[8] Andrea Kulkarni, *Bias in AI and Machine Learning: Sources and Solutions*, LEXALYTICS BLOG https://www.lexalytics.com/lexablog/bias-in-ai-machine-learning#:~:text=There%20are%20two%20types%20of,certain%20expectations%20in%20our%20thinking.

[9] ibid.

[10] Supra note 8.

[11] Scott Fulton, *What is Bias in AI really, and Why Can't AI Neutralize It* , ZD NET (June 15, 2022) https://www.zdnet.com/article/what-is-bias-in-ai-really-and-why-cant-ai-neutralize-it/

[12] Ani Aggarwal, *AI and Fake News in Social Media*, MEDIUM (June 15, 2022) https://medium.com/nerd-for-tech/ai-and-fake-news-in-social-media-88fedfc11ad

[13] Emma Llansó, Joris van Hoboken, Paddy Leerssen, Jaron Harambam, *Artificial Intelligence, Content Moderation, and Freedom of Expression*, TRANSATLANTIC WORKING GROUP (June 15, 2022) https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf

teaching AI about skin tones, then the AI will not classify the nudity of a person with a darker skin tone as harmful content. Another example of underrepresented in datasets creating efficient AI is "*if these datasets do not include examples of speech in different languages and from different groups or communities, the resulting tools will not be equipped to parse these groups' communication.*"[14]

Other than underrepresentation in datasets, there is another problem of there being insufficient data to teach the AI. In 2019, a man live-streamed a video of him opening fire in a mosque in New Zealand and killing 51 people. The AI was not able to detect this gruesome video and remove it from the internet because it did not know that videos of someone killing other people is harmful due to the fact that the data sets did not include such videos. Facebook's chief artificial-intelligence scientist justified AI's mistake by saying that "Thankfully, we don't have a lot of examples of real people shooting other people."[15]

AI can be taught or trained in two different ways. The first method is called supervised learning. In this method, the AI learns through looking at human annotated data which is in the form of bulk of texts or news articles accompanied by human commentary. In this process, the AI becomes capable of sorting out even new data that it has not seen before.[16] For example, a social media platform building an algorithm to identify hate speech is given a collection of posts classified in two categories – 'hate speech' and 'not hate speech'. The annotations or labels given to the content by the human act as guides for the AI in identifying and differentiating videos.[17] The problem with this model of training is that annotating datasets is a very time and cost consuming process and hence most of the annotated datasets are quite old and thereby not a correct representation of the current world scenario.[18] Moreover the process of generating the dataset and having humans label them could also lead to introducing biases that the human may have, into the algorithm.[19]

The second method used to train AI is called unsupervised learning. In this, the AI is presented with raw data without any labels or annotations, and the AI is made to observe, read into and

[14] ibid.
[15] Ben Dickson, *Human Help Wanted: Why AI is Terrible at Content Moderation,* PC MAGAZINE DIGITAL, (June 15, 2022) https://www.pcmag.com/opinions/human-help-wanted-why-ai-is-terrible-at-content-moderation
[16] Supra note 12.
[17] Supra note 13. Pg. 4.
[18] Supra note 12.
[19] *Use of AI in Online Content Moderation*, Report on behalf of Ofcom, CAMBRIDGE CONSULTANTS, Page 41, 2019 https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf

uncover the natural patterns in the data.[20] One of the biggest advantages of unsupervised learning is that since the data does not need to be labelled, the datasets can be much larger. This is also helpful for creating more up-to-date datasets thereby allowing a wider range of information that's relevant to the real world. Additionally, since unsupervised learning is free annotations, the chances of human bias entering the algorithm are comparatively less if not none.[21] Human biases may still creep in because the data given to the AI is still picked and compiled by humans, but removing the annotation requirement definitely helps in significantly reducing the bias. Other than this, auditing also becomes a problem because of the insanely large unsupervised datasets. A unique feature of this system is that AI begins to create their own intuitions.[22] The AI finds its own way of sorting out data when trained using this method which is both an advantage as well as a disadvantage. When unsupervised AI systems are using in unification with human moderators, they can detect fake news more efficiently than the humans.[23] But at the same time, since the reasoning of the AI system in reaching a particular decision is not known to the human moderator, it gets difficult to judge or predict AI's actions and hence hard to control them. But, in terms of bias, unsupervised learning is definitely a better option than supervised learning.

Even while using the unsupervised training method, special care should be taken to ensure that the dataset is not biased, and is inclusive of every community in the society. This is the biggest problem that arises in terms of AI being biased. "*A biased dataset can be unrepresentative of society by over or under-representing certain identities in a particular context.*"[24] Even if the dataset is accurate, it might be representative of an unjust society and hence again brings the societal bias into the algorithm and the way it functions. For example, in a study, it was found that tweets written in African American English (used by Black Americans) was twice as likely to get flagged as offensive.[25] But the point that needs to be noted is that what is considered offensive differs from place to place. Slurs that are considered normal in certain areas may be taken as offensive in others. But, AI systems lack the ability to understand such nuances and complexities. Context, which is the most important factor in such situations, is beyond the scope of AI and thereby goes on to amplify bias by discriminating against a particular section

[20] Supra note 12.
[21] Supra note 12.
[22] Supra note 12.
[23] Supra note 12.
[24] Supra note 7.
[25] Merlyna Lim, Ghadah Alrasheed, *Beyond a Technical Bug: Biased Algorithms and Moderation are Censoring Activists on Social Media*, THE CONVERSATION, (June 15, 2022) https://theconversation.com/beyond-a-technical-bug-biased-algorithms-and-moderation-are-censoring-activists-on-social-media-160669

of people. Algorithmic bias of this kind can very negatively affect those who are already at-risk due to being underrepresented or misrepresented.

Bias can creep in when the purpose and constraints of an AI model are defined.[26] AI can lead to technically incorrect and discriminatory conclusions for certain section of the population if the quality of data used in training is biased or unjust.[27] Another way in which bias can enter the AI system is when the algorithm is not used properly. If the algorithm is not evaluated well, then irrespective of the quality of data used, the outcome could be biased. If the AI system was developed for being used in a particular situation or context or for a certain section of people, and is being used for some other purpose, then bias is inevitable.[28] This is because the AI would not be able to capture the change in society and values and hence cannot adapt itself to give a right conclusion even when the social context changes. A recent example of bias in AI was seen with Twitter's image cropping algorithm. The algorithm was designed in a manner that when one uploads a picture on Twitter, it crops the image focusing on a human face and displays it as the preview. But, it was observed that the algorithm always only cropped the image of a white person, preferably male when the picture had people of different race and genders.[29] A number of photos, including those of dogs and cartoons were tested but all of them showed preferential treatment against the colored people. Twitter had apologized and taken down the racist algorithm but this example goes on to show that even a social media giant like twitter, who has access to all sorts of data in the world used an algorithm that is so biased.[30]

AI bias is not something that should be taken lightly. Since there is an excessive demand for the use of AI in almost every field of life, it becomes absolutely important to understand the kind of negative impact biased algorithms can have on us individuals and the society as a whole. Biased AI can affect the well-being of people by treating them unequally, being derogatory or offensive to them and questioning their existence.[31] AI systems can encroach upon civil liberties and rights of people by reinforcing existing prejudices of the society. If bias in AI is not checked, then with the growing usage of AI in today's world, it would only lead to solidifying and crystalizing the inherent societal bias and discrimination. Biased AI affects not only the marginalized groups in the society about whom there is less data available, but also

[26] Supra note 7.
[27] Supra note 7.
[28] Supra note 7.
[29] Supra note 8.
[30] Supra note 8.
[31] Supra note 12.

affects all of us. Because it reduces the potential of AI for business and society by increasing mistrust on AI in general.[32]

Having said that, removing bias from AI is an extremely difficult and almost impossible task. The decisions that AI makes can get very unpredictable and opaque. This makes it a challenge to analyze and scrutinize the performance of AI and thereby poses to be a hurdle in AI's or its makers accountability when it causes harm to someone. This problem is generally referred to as the 'black box' problem, because the AI with time begins to develop its own logic by identifying patterns that are difficult to explain normally.[33] Even if two neutral systems are trained using the same dataset, they may both end up learning differently resulting in them being inconsistent with the detection of harmful content across platforms.[34]

It is very important to find ways to mitigate this bias. People designing AI shall be mindful that the algorithms shall be fair. Since one of the primary sources of bias in AI is the data used to train the AI systems, it should be ensured that the data is not biased or unjust.[35] It is not practically possible to completely erase bias out of the data because the data is a mere representation of our society which unfortunately is biased and unfair. But certain measures can be taken to at least try and prevent such biases entering the AI systems. Data can be processed beforehand, teaching the AI system about fairness and sensitive concepts like gender, race, sex etc. during the training process itself.[36] Special attention should be paid to the fact that the AI does not use historical data so as to avoid pre-conceived judgments because if that happens the mistakes of the past will only repeat themselves.[37] The data used shall also be latest and shall cover all parts of the society to avoid discrimination against anyone. Another step in that direction would be to include more and more people from different backgrounds in the process of collecting or assembling data for training AI systems.[38] Organizations using AI should also try and be more transparent to the public or its users for them to know the logic used by the AI in reaching a conclusion. This will also increase accountability and create a

[32] James Maynika, Jake Silberg, Brittany Presten, *What Do We Do About The Biases in AI?*, Harv. Business Rev., 2019 https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai
[33] *The Right to Privacy in the Digital Age: Report on the United Nations High Commissioner for Human Rights*, Human Rights Council, 48th Session, 2021
[34] Supra note 19.
[35] Steve Nouri, *The Role of Bias in Artificial Intelligence*, FORBES TECHNOLOGY COUNCIL (June 15, 2022) https://www.forbes.com/sites/forbestechcouncil/2021/02/04/the-role-of-bias-in-artificial-intelligence/?sh=2b15dc4e579d
[36] Supra note 32.
[37] Theodora Lau, Uday Akkaraju, *When Algorithms Decide Whose Voices Will Be Heard*, Harv. Business Law Rev., 2019 https://hbr.org/2019/11/when-algorithms-decide-whose-voice-will-be-heard
[38] Supra note 25.

sense of trust. The AI systems should be subjected to regular audits and checks so as to effectively identify any patterns of bias. AI should be fair, safe, non-discriminatory and in correspondence with human rights. AI should not be trusted with the complete responsibility of moderating content online, but it shall be used along with human moderators as additional support so as to ensure human oversight which might help in reducing bias.