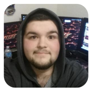

3-1 Discussion: Hidden Bias in AI

Listen

Back to Topic

Hidden Bias in AI

Nick Hahn posted Jul 13, 2023 11:48 AM ★ Subscribed

Bias Example

As highlighted in one of the articles, AI models may possess bias due to the data used to train them. This shows in various results such as a husband and wife receiving different credit limit results from the same application for an Apple Card (Heilweil, 2020). As the article points out however, without transparent access to the training data we cannot make the determination one way or the other. Without evidence to back the claim of discrimination, it is just as likely that there were valid reasons for the wife receiving a lower limit than the husband. This matter is complicated by the fact that transparency in machine learning applications is not as straightforward as it may seem (Kopf, 2019).

Bias Mitigation

Efforts to remove or limit biases include critical thought precursors such as defining the business case for the model, thorough understanding of the training data, accounting for the potential end-users, and ensuring diversity of thought and perspective regarding the development team (Janet, 2021).

Application

By employing heavy doses of critical thought, careful selection of unbiased data, and careful definition of the business case by a thought and perspective-diverse team, the resulting model can better approximate reality and produce valuable results for the specific endeavor. Additionally, by utilizing such tools and having a robust process for development of the model, the results of the model can better stand up to scrutiny, regardless of human-biased output desires. Perhaps the husband and wife in the example may receive more similar credit limits, but its is also possible that the credit limit disparity may remain, proven out by a critiqued training data set and a rigorously-developed model.

References

Heilweil, R. (2020, February 18). Why algorithms can be racist and sexist. Vox. https://www.vox.com/recode/2020/2/18/21121286/algorithms-bias-discrimination-facial-recognition-transparency

Janet. (2021, May 14). Overcome and prevent bias in AI. Figure Eight Federal. https://f8federal.com/overcome-and-prevent-ai-bias/

Kopf, D. (2019, November 15). Goldman Sachs' misguided world cup predictions could provide clues to the Apple card controversy. Quartz. https://qz.com/1748321/the-role-of-goldman-sachs-algorithms-in-the-apple-credit-card-scandal

Reply to Thread

Filter by: All Posts | Clear filters Show: Threaded

Aakash Thapa
July 14 at 9:52 AM

Hello Nick,

I thoroughly enjoyed reading your post. You mentioned one of the solutions is a thorough understanding of the training data. I concur with your viewpoint that a comprehensive understanding of the training data is crucial for detecting and addressing biases that may be present and could impact the AI model. It is important to assess the representativeness, diversity, and potential biases within the training dataset.

The alternative to this can be implementing a comprehensive data auditing and preprocessing step that specifically aims to identify and mitigate biases. This can involve techniques such as data sampling, data augmentation, or even using synthetic data generation methods to create a more diverse and balanced training dataset. Additionally, incorporating external datasets or diverse perspectives can help counteract biases present in the original training data (Caliskan, 2017).

Reference:

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

Reply

Nick Hahn
July 17 at 10:49 AM

Hi Aakash,

I like your alternative suggestion; reminds me of a unit-testing development approach. I could see that being a robust way to build a system with requirements for transparency built into the design model.

Best regards

Nick

Reply

Timothy Alexander
July 15 at 8:42 PM

Great write-up, Nick! You make a great point about knowing whether or not there is bias in an AI application. You're absolutely right that you can't know about the bias unless you look at the training data. I would suggest that as a staring point, for sure, because that's honestly the most likely place that you're going to find bias coming into an AI system. However, if you review the data and don't find any reason for the system to have become biased, then you have to dig deeper and look at the algorithms themselves to see if the algorithm may have been written in such a way that it could be misinterpreting the data it was fed.

Reply

Jordan Goyena-Segarra
July 16 at 5:38 PM

Hey Nic,

your bias mitigation is a really good start but I believe more is necessary. Narrowing the problem as you have said is a great way to start and anything to limit the bias initially will help. That being said I believe iterating on the model and retraining with new information after showing it to tester will help more.

~Jordan

References

Janet. (2021). Overcome and prevent bias in AI. *Figure Eight Federal*. https://f8federal.com/overcome-and-prevent-ai-bias/

Reply

Joel De Alba
August 4 at 12:04 PM

Your example of bias in AI models and the potential credit limit disparity for a husband and wife is a pertinent one, as it highlights the real-world implications of biased algorithms. It's essential to address these biases to ensure fair and ethical AI applications. Your proposed solutions for bias mitigation, such as defining the business case, understanding training data, considering end-users, and promoting diversity in the development team, are indeed valuable steps.

To complement your suggestions, I would propose the incorporation of fairness-aware algorithms and interpretability techniques. Fairness-aware algorithms aim to reduce or eliminate bias by explicitly considering fairness constraints during model training (Kusner et al., 2017). Techniques like adversarial learning and re-weighting of training samples can be used to address bias in the data and ensure equitable outcomes.

Moreover, interpretability techniques can enhance transparency, helping stakeholders understand the decision-making process of AI models. By using methods like LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations), we can gain insights into how the model arrives at specific decisions, which is crucial for detecting and addressing hidden biases (Lundberg & Lee, 2017).

By incorporating fairness-aware algorithms and interpretability techniques in addition to your proposed solutions, AI models can become more accountable, transparent, and fair, promoting trust and mitigating potential biases.

Sources

Kopf, D. (2019, November 15). Goldman Sachs' misguided world cup predictions could provide clues to the Apple card controversy. Quartz. https://qz.com/1748321/the-role-of-goldman-sachs-algorithms-in-the-apple-credit-card-scandal

Heilweil, R. (2020, February 18). Why algorithms can be racist and sexist. Vox. https://www.vox.com/recode/2020/2/18/21121286/algorithms-bias-discrimination-facial-recognition-transparency

Najibi, A. (2020, October 24). *Racial discrimination in face recognition technology.* Science in the News. https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

Reply

‹ 1 / 1 ›

Reply to Thread

Reflect in ePortfolio

Activity Details

Well done! You have contributed to the discussion

Available on Jul 8, 2023 10:59 PM. **Submission restricted before availability starts.**

Assessment

0 / 30 **F**

Discussion Rubric: Undergraduate

Last Visited Aug 9, 2023 3:59 PM