Destin Saba
ID: 30249241
September 15, 2024
ENSF611 - Assignment 1

1. What does AI ethics mean to you? (1 mark)

To me, AI ethics are fundamental human aspects that should govern the development and use of any AI. AI ethics should ensure that all AI is fair and considerate of human rights. Under the umbrella term ‘fair’, I think ethics should specifically consider:

- Transparency of AI models (their decision-making should be easier to understand)
- Data privacy (the data that goes into a model should be done so consensually by its originator)
- Removal of bias (Biases in input data should be screened out so that AI does not carry those same biases and cause harm)

2. Please watch this YouTube video (same one as mentioned in class) and answer the following questions:

a. How did you feel after watching the video? (1 mark)

I was in a state of reflection, and began to wonder what biases I have had in my past that may have shown up in my work. While no specific example came to mind, I am sure I have had some - and I will therefore be perceptive of my biases in my future work.

b. List three things that you learned from the video (3 marks)

1. Just a single biased dataset can end up informing many software programs all over the world, very quickly, since a lot of developers seem to use similar public datasets.
2. That half of the adults in the US have their faces in facial recognition networks, and that police networks can look at these networks unregulated (at least as of 2017 when the TED talk was released).
3. Some judges use machine learning generated risk scores to help determine how long an individual will spend in prison.

3. In the video, Joy lists many different examples of algorithmic bias in society. Please find an example of algorithmic bias in society (you can use one mentioned in the video or find your own) to answer the following questions:

Destin Saba
ID: 30249241
September 15, 2024
ENSF611 - Assignment 1

a. Online link to information about your example (1 mark)

https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

b. Summarize the example you have chosen (2 marks)

I chose an example Joy mentioned in her TED talk that really caught my attention: machine learning algorithms are being used to conduct risk assessments on a defendant's profile to inform a judge's sentencing decision.

The algorithm, which is based on historical crime data, will provide the judge a number that indicates the likelihood that the defendant will commit another crime.

One of the intents of this risk assessment was to rely less on an individual Judge's gut and biases, and instead rely on "impartial" data. However, there are strong biases in historical data against those coming from low-income or minority communities, which will result in harsher sentencing towards those communities. This can become a dangerous cycle that will only produce more biased data, which will then be used to train even more biased models in the future.

c. Based on the five different types of bias in the notes, what type of bias does this example represent? Justify your answer (2 marks)

The machine learning model in this example has group attribution bias, where "assumptions are made about individuals based on the group they belong to". In this example, due to a correlation in historical data between low-income / minority communities and crime rates, the risk score for individuals from those communities will be harsher. Neither being from a minority nor coming from a low-income community should be metrics used in determining an individual's likelihood to commit a crime.

Since this model is predicting an individual's behavior considering the group they come from, this is an example of group attribution bias.

d. Why did you pick this example? (2 marks)

I chose this example since it was the most shocking to me when I first heard it. Times change, laws change, and society changes so quickly, so it didn't make sense to me when I heard that machine learning models which rely on historic, outdated data are being used to make sentencing decisions today.

Destin Saba
ID: 30249241
September 15, 2024
ENSF611 - Assignment 1

I also believe that a part of the sentencing decision should in fact come from the individual Judge's gut. I think that there are many human factors that can't (at least not yet) be completely captured with the historic data that is available.

e. Describe one way that you could fix this issue (2 marks)

If removing the algorithm all together is not an option, I would suggest re-training the models with a dataset that does not contain group demographics that could lead to group attribution bias in the model. Only data that is impartial to everyone should be used, like the type of crime committed, if it was a second offense, or the age of the defendant.

4. In class, we discussed how to properly use generative AI and some of the ethical dilemmas.

a. What are two different ethical ways that you could use generative AI for this course? (2 marks)

1. Generative AI could be used to generate practice questions on a topic that I am studying.
2. Generative AI could be used to brainstorm ideas in a conversational manner relating to any topic discussed in class.

b. Do you think software companies should pay for any copyrighted materials that are used in their training models? Why or why not? (2 marks)

I think that software companies should pay for copyrighted materials that are used in their models. Since the software company will be extracting value from their model, and therefore also the copyrighted materials, the originator of the copyrighted materials should be compensated fairly.
Additionally, by compensating the creators of training data, creators will still have an incentive to make high-quality work, and this will also result in better quality models.

c. Pick one of the other ethical issues surrounding generative AI that were discussed in class. Describe one way that companies could address this issue. Why did you pick this issue? (2 marks)

From the lecture slides, we briefly discussed an example of how "a loan approval system may unfairly judge applicants based on their gender and marital status". This

could happen if the algorithm used was trained on historical data that contains outdated social norms and stereotypes (Pazzanese, 2023).

I think one way companies could address this issue, along with any other AI ethics related issue, is to embed an AI ethical review into the software development process. A thorough dataset review should be completed before any model development begins, and the model should be tested for any lingering biases once complete. In this specific example, factors such as marital status and gender should not have been considered at all. I think relevant data would include things like household income, years of employment, industry of employer, and credit score.

References:

1. OpenAI. (2024). *ChatGPT (Version 4) [AI model]*. OpenAI. https://chat.openai.com
2. Dawson, L. (2024). *Introduction to AI Ethics* [PPT]. ENSF 611., University of Calgary.
3. Hao, K. (2019, January 21). *AI is sending people to jail—and getting it wrong*. MIT Technology Review. https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/
4. Buolamwini, J. (2020, September 10). *How I'm fighting bias in algorithms* [Video]. TED. YouTube. https://www.youtube.com/watch?v=UG_X_7g63rY
5. Pazzanese, C. (2020, October 7). *Great promise but potential for peril*. Harvard Gazette. https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/