Data Ethics In-Class Activity (May 23)

Today we will discuss two of the Ethics subtopics, Data Privacy and Validity. Refer to the corresponding articles listed in the previous activity document. There is no need to submit any written report about the topics or about the discussion. Be sure to stay in your breakout room even if you finish your group discussion early, as there is a class discussion at the end of the session.

Data Privacy

Your job is to discuss the Data Privacy topic first within your breakout group (for approximately 30 minutes) and then join the main room for a full class discussion of the topic.

Within your group:

1. Who read any of the articles about Data Privacy?
2. Which article(s) did you read?
3. Each person who read any of the articles should take a few minutes to give an overview of the article, specifically:
    a. Who wrote the article, what is their role or point of view?
    b. What are the main points of the article?
    c. What are the strong points of the article?
    d. What are the weak points (if any)?
    e. What did you learn from it or take away from it?
4. All others in the group then should discuss and ask questions about the article.
5. Be ready to discuss the following questions with the full class
    a. What is the GDPR?
    b. GDPR is a European effort, how does it relate to the USA?
    c. How might a Data Engineer be involved in GDPR compliance?
    d. Discuss the following questions:
        i. Popups everywhere. It's annoying and the average internet user has no idea how to control/configure data privacy consent, so they just agree to everything.
        ii. Companies are scared, so they are spending bajillions protecting themselves. Bajillions that could be spent on things that actually benefit customers.
        iii. The whole thing is toothless. Only Ireland can bring an actual judgment, and they are in the pocket of big tech. So there have not been many significant cases or judgements so far.
        iv. It requires private data to be transparent and easily accessible by the users, and that makes it easier for hackers to obtain private data by impersonating users.

Validity

Discuss the Validity topic first within your breakout group (for approximately 30 minutes) and then join the main room for a full class discussion.

Within your group:

6. Who read articles about Validity?
7. Which article(s) did you read?
8. For each article read by at least one person in the group:
    a. Who wrote the article, what is their role or point of view?
    b. What are the main points of the article?
    c. What are the strong points of the article?
    d. What are the weak points (if any)?
    e. What did you learn from it or take away from it?
9. Discuss the following questions:
    a. The articles list many problems with data validity. Which of these problems could be helped by a Data Engineering approach?
    b. What specifically could/should a Data Engineer do to address the challenges listed in these articles?

Submit

Create a copy of this document (or create a new document if you prefer), and use it to answer the following question.

For each of the four major areas of Data Ethics, mention a situation that you have experienced that involved the corresponding area of Data Ethics. Say whether or not (in your opinion) the issue was handled satisfactorily. Finally, state how you might improve the handling of Data Ethics in similar situations in the future.

Use the in-class assignment submission form to submit your response(s).

Ownership

In terms of data ownership, what comes to mind would be how social platforms manage your data and the fact that you have no control over it once you have given it to them. Even if they offer to delete all your information, there is a good chance that they keep the data points in an obfuscated form to avoid any legality issues, if they even do that.

Privacy

When the internet was a fairly new thing in the early 2000s(at least for mainstream use) most ISPs in Sweden advertised that they save no data on their users. Probably because music piracy and piracy in general as I remember it was fairly common. But with the pressure from the music industry on the whole hunt for the pirate bay owners legislators made it so that ISPs had to store a certain amount of logs on users. I don't know if I think this was the correct choice in this case. Everyone had to give up part of their privacy for corporate concerns. I think that at the time the music industry was charging too much for their music and today with streaming we have found a middle ground where the music is cheap enough to get everything for a monthly fee from a provider over the hassle of piracy. This is how piracy issues should be solved. Most people will choose good if it is a reasonable option, and tend to gravitate towards bad if good is put on a too high shelf.

Validity

I personally have never been in a situation where I had issues or experienced major validity concerns in an ethics related situation. But I could imagine the consequences of invalid or data rot affecting outcomes of people or businesses.

Fairness

In the CS ethics class I chose to do research on AI in medicine and found some great articles on how medical diagnostic tools were under diagnosing minorities based on historical data when there was conflicting data whether or not they fell into that category or if it was just picked up because historically that population had been commonly overlooked. I think in our discussion Sai said something important, that with the statistical biases surfacing and coming out of the woodworks, it becomes harder to marginalize and overlook these biases since the numbers are better evidence than a couple of anecdotal "I was mistreated" stories. As these biases gets discovered we can start working towards more fairness in our data driven world.