Ethical Challenges in Data Science

RCR Orientation for MIDSters

Olanrewaju Michael Akande

Assistant Professor of the Practice
Social Science Research Institute and the Department of Statistical Science
Duke University

Outline

Ethics in data science

- As a data scientist, you will often be faced with the moral problems associated with data and the use of data perhaps more often than you might like.
- These moral/ethical problems will come in so many different forms, across the entire pipeline of data science.
- From data collection, to modeling and analysis, to the deployment of algorithms, to the interpretation of results, to protecting data from hackers, to day-to-day maintenance of technology, and so much more!
- Ethical issues in data science often arise in many more serious, painful and costly scenarios than many of you may have heard/thought of.

Really???

Here are a few examples of those scenarios:

- Dissecting racial bias in an algorithm used to manage the health of populations
- L.A. is suing IBM for illegally gathering and selling user data through its Weather Channel app
- Amazon scraps secret AI recruiting tool that showed bias against women
- Goldman faces probe after entrepreneur claims gender bias in Apple Card algorithm
- Critics say a childhood asthma study unethically withheld care—and see a troubling trend
- Robodebt was a fiasco with a cost we have yet to fully appreciate (DHS in Australia)
- Timnit Gebru's actual paper may explain why Google ejected her (for context, Gebru co-authored this paper: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?)
- Evidence of Fraud in an Influential Field Experiment About Dishonesty

So many questions...

Many of the ethical questions in these examples and the many scenarios you will encounter throughout your career often revolve around the following broader questions:

- Who owns data?
- What does user consent really mean when collecting data?
- Who has the right to access data?
- Under what conditions can one be provided access to data?
- What is the plan to protect and secure data?
- What does fairness even mean?

Just tell me what to do...

OK what then?

- There are no universally accepted or straightforward responses to these questions.
- In fact, the purpose of our session today is not to "solve" a case study or to arrive at a right/wrong answer.
- The goal today is not to"teach" you ethics.
- Rather, the goal is to get you to engage in authentic dialogue with each other and myself, about ethical issues that can arise across the different facets of data science, and considerations for how one might move forward.
- The following case studies are therefore provided as models for potential complexities and how to think critically about these issues.

Case Study I: the ethics of using hacked data

The reading can be found here: The Ethics of Using Hacked Data: Patreon's Data Hack and Academic Data Standards

You will be split into small Zoom breakout rooms.

Take 10 minutes to read the material (scan through to get a general idea if you need more time) and another 10 minutes to discuss the study and the questions.

Then, reconvene and engage with the whole group.

Case Study II: rigorous model assessment or a happy supervisor?

The reading can be found on Sakai. Go to the RCR Resources folder, then

→ Engr and Nat Sci - Ethical Case Studies

→ Math_Statistical Science Ethical Cases

→ IncorrectAnalysis.Case10.pdf.

You will be split into small Zoom breakout rooms.

Take 5 minutes to read the material and another 10 minutes to discuss the study and the questions.

Then, reconvene and engage with the whole group.

Case Study III: who can change Proprietary Source Code?

The reading can be found on Sakai. Go to the RCR Resources folder, then

→ Engr and Nat Sci - Ethical Case Studies

→ Math_Statistical Science Ethical Cases

→ Comp Sci Case Study - Proprietory Source Code.pdf

You will be split into small Zoom breakout rooms.

Take 5 minutes to read the material and another 10 minutes to discuss the study and the questions.

Then, reconvene and engage with the whole group.

Is that it?

- Clearly, the conversation around ethics in data science needs to be an ongoing one.
- You will revisit this many times during your time here at Duke, particularly within MIDS.
- During the spring semester, the Unifying Data course will touch on machine learning bias in particular and provide some recommendations for you.
- In your second year, the Data Science Ethics course will dive deeper into your social responsibility as data scientists.
- Then of course, many of you may encounter similar questions during Capstones.
- Finally, always probe and consider the moral implications of your work as you go forward in your career.

By the way...

Take a quick peek at the following links:

- A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle
- Doing Data Science: A Framework and Case Study
- The Council for Big Data, Ethics, and Society Case Studies

Questions?