HUY DAI NGUYEN

MIS 2502-Data Analytics

Prof. Shunting Wang

**DATA SCIENCE ETHICS**

“Data” has been the most popular buzzword in business and STEM industry. There’s also another buzzword that is on the rise as well: “privacy issues”. For the past 2 years, we have seen many cases of data breach or misuse that were reported to the public. As more users are becoming terrified and distrustful of tech companies managing their personal data, data scientists are coming together to understand and establish its code of ethics, also known as “data science ethics”.

So what are the ethics of data science? To understand data science ethics, we should have an overview of what is ethics. Ethics is a system of moral principles that concerns with the “good” and “bad” defined by individuals and society. It’s also an important determinant of a person or organization’s reputation. Those who follow ethical practices are well-respected, while those who violate them are controversial and lowly regarded by others. Based on this definition, data science ethics refer to a branch of ethics that evaluates moral issues related to data, algorithms and corresponding data practices. Through ethical data collection, analysis and implementation, data scientists can generate solutions that are “morally” good for both business and users.

With our data constantly collected by companies every day, it’s a necessity to know why data science ethics matter so much to us and to data scientists. For users, it’s a problem of privacy infringement. If there’s a security breach that allows private data to be out in the open, a user can be subjected to identity theft, manipulation or real-life consequences. Consider Equifax data breach in 2017 in which 145 million accounts were compromised. Hackers gained access to names, SSN, address, driver’s licenses etc., which allowed them to falsify information to set up credit cards, loans under the victims’ names. For data scientists, they need to establish data ethics due to the growing distrust from its customers. Lack of transparency and cooperation can lead to boycotts or bad press that would be damaging to the company.

There are 4 topics concerning data science ethics: ownership, bias, transparency and consent. Ownership, referring to ethics of practices, asks “who should be the owner of the data? The customers or the company?”. Bias, referring to ethics of algorithm, determines if there’s discrimination or favoritism within the data used for analysis (which sometimes is not a bad thing). Transparency, referring to ethics of data, involves how the data collection process and usage will be disclosed and how the data will be protected from unwanted external third parties. Finally, consent, referring to ethics of practices, determines how companies can get your approval to use your data.

With the growing number of data breaches, there have been implementations on possible solutions to data misuse and privacy infringement. In Europe, government has taken actions towards data control through the EU General Data Protection Regulation (GDPR). The law, in addition to putting restrictions on data collection from companies, states user rights on their data, including the right to have your data deleted, to change your own data, to be notified when a company leaks your data etc. More importantly, people need to raise their awareness and to facilitate more discussions regarding data ethics. In fact, data scientists and leaders are coming together to as a community to establish guidelines and the “Hippocratic Oath” for data scientists. One thing is for sure: progress is being made. And that’s one breach at a time.