Ethics Discussion #1

**Q 1.1:**

The training data that the AI is trained on could be biased on what data is being feed in or what data is being looked at for the AI to make its decisions. For the case of McDonalds in Chicago, there was dirty data/not clear enough data to make a link from crime to him for a human, but the AI took leaps to jump to the conclusion.

**Q 1.2:**

It’s hard to prove that AI’s conclusion is not in line with the data because for one, it is in the sense that the paths/jumps are correct. The AI may not have enough or clean enough data to actually make a useful conclusion. Also, many of the victims of misclassification are not in positions to go after big tech companies or police stations that use them.

**Q 2.1:**

AI is used to monitor content that is posted to the social media site in the form of “reading comments”, “watching videos”, and “looking at photos”. That is the main use of AI that the everyday person will see but then there is a “secret algorithm” that is tailored to you to serve up content right to your home screen on the social media.

**Q 2.2:**

In the “Enshitification” article, it boils down to how creators of content on the platform will make loopholes around the use of “bad” words on the site. For example, “shit”, “death”, “suicide”, “terrorism”, etc.

**Q 3.1:**

Bender/Hanna’s current harms include but not limited to, discrimination in housing, wrongful arrests, and deep-fake pornography. “Safe.ai”’s future harms include bioterrorism, a Military AI arms race, and unleashing an AI that will destroy humanity.

**Q 3.2:**

They would suggest that AI currently does not have enough power to surpass human intelligence. In the future maybe there is a day when AI does surpass human intelligence but there needs to be checks and balances in place when that day arrives

During class:

Basis

How can a computer system lead to racist outcomes even if it doesn’t explicitly code for race?

* Training Data 🡪 Racist? 🡪 Racist AI
  + Supervised learning; learning from examples.
  + Face ID people of color
  + All training data have bins
  + Computational photography; bad contrast
  + What’s in the data?
    - Open source??
    - Teaching? provenance
* Testing team 🡪 google maps “Malcolm 10 Boulevard”
* Goal being set
* Most systems give an answer
  + Not enough information to answer but gives the closest answer anyway
* Proxies for basis
  + E.g., Zip code; Dunkin Donuts vs Starbucks; name

Prop: It is impossible to avoid bias in a human creation

Automating Data

Social media

* Detect violations against TOS for posts
  + Appeals for TOS
* Cost cutting
* Machines don’t need to sleep
* Consistency
* Serving content 🡪 recommendation algorithms
  + Long tail graph
  + Data Fusion