Ethics Discussion #5

**Q1.1**

A big issue could include diversity of data because as the paper states just because you have a large dataset does not equal a large amount of diversity in that dataset. There are plenty of other issues like biases of the people who put the training data together for race, sexuality, gender, etc. etc.

**Q1.2**

Documentation debt is when the datasets are either or both undocumented or too large to document after the fact. Undocumented is referring to the characteristics of the data that could be very useful for training.

**Q1.3**

The authors talk about how the models perverse the form but not the meaning behind any of the output because of the lack of meaning in the training datasets.

**Q1.4**

On page 617, right before section 6.2 we see what the authors are getting at where the “stochastic parrot” is just a language model that is stitching together sequences of linguistic forms from its training data according to the probability that that sequence will happen again. Our own “linguistic competence” does trick us into thinking that the LM has meaning behind the words they say because we are taking the implicit meaning behind the output of the LM.

**Q2.1**

First off, they talk about how every single news article was going to be about a large language model. Along with the explosion of AI over the last 2 years where it’s gone from the niche aspects of computer scientists and researchers to a worldwide audience.