**Cataloguing Bias in GPT-3**

Kieran Ahn

Mentor: Dr. Korpusik

Abstract:

GPT-3 is an incredibly powerful NLP model that has taken the NLP world by storm. However, it possesses the glaring flaw of bias: due to its training datasets being sourced from the internet, they contain biased and prejudiced text which shows through in some of GPT-3’s text output. In order to eliminate this bias in whatever way possible, it is necessary to first understand it and figure out what its biases are in the first place. Therefore, I propose an exhaustive analysis and evaluation of GPT-3’s biases towards a wide variety of people groups in the hopes of creating a reference for other research to use when investigating ways of combating bias in GPT-3.

**Introduction:**

Talking with computers has never been more advanced than today. We have progressed from the primitive language of zeroes and ones through punch cards to languages like Assembly, and then proceeded to develop even more advanced technologies, branching out into languages like C++, Java, and Python. However, those efforts, no matter how high the abstraction, were limited to the one-sided and inflexibly logical rules of programming algorithms with pre-written rules and a limited range of outputs. We could tell computers what we wanted, but they could not speak back—not like the way we can. Now, however, speaking technology is everywhere: Siri, Alexa, Cortana, and other virtual assistants are ubiquitous in modern-day households. Yet those helpers are still obviously computers; their responses are rigid, predictable, and pre-set. However, on the cutting edge of artificial intelligence technology, OpenAI’s new API powered by GPT-3 is making waves as one of the most advanced natural language processing algorithms to date, and certainly the largest (Brown et al. 8).

GPT-3 works through language modeling (Brown et al. 11), which is essentially determining the probability of certain words appearing in a sentence. For instance, you, reader, may perform the same operation when I write: “the dog goes…” or “two plus two equals...” and other prompts like that, only GPT-3 does so having trained on thousands of books and data pulled directly from the internet (with all of the consequences that would entail) (Brown et al. 9). It should require no stating that the internet is a volatile, censor-free, wild west where anybody can spew hateful, prejudiced ideas and avoid reprisal (and in many cases, gain the opposite). Biased data like that is a minority, but is nonetheless a part of what GPT-3 uses to help generate its output. In essence, that process of determining which words are likely to come one after another in a sentence is a method of quantifying the zeitgeist of our culture. The frequency with which GPT-3 generates biased text can show us how often prejudiced ideas are articulated. Preliminary research done by the OpenAI research team behind the initial GPT-3 paper focused on gender, race, and religion (Brown et al. 36). Possible ways of expanding that research include broadening the categories tested, such as testing other topics such as political affiliation, sexual orientation, and so forth. Another addition would be to deepen the testing on the pre-existing topics, using the advantages of GPT-3’s algorithm to quantify and summarize additional opinions on the topics the OpenAI team have already presented.

**Background:**

The topic of uncovering bias in NLP systems is not a new one. In 2019, Sheng et al. authored a paper dealing with the biases inherent to two NLP systems: GPT-2, the predecessor to GPT-3, and a language model used by Google. This paper covered sexuality (gay & straight), race (black & white), and gender (man & woman). Sap et al. created the idea of “social bias frames”, a model representing the ways “in which people project social biases and stereotypes onto others” (1) in terms that NLP systems would be able to understand, with the intent that future researchers would use them as a way of mitigating bias in future NLP models. They also highlighted the dangers of inherent bias in the dialects present in the datasets used to train NLP algorithms: using a dataset which predominantly or mostly contains a single dialect could result in the model inheriting the speakers of that dialect’s biases, and also under-represent minority dialects. Moreover, GPT-3 is clearly biased. The OpenAI team states it plainly in their paper (Brown et al. 26) and their website (Brockman et al.), and users have already identified harmful biases in the text generated by GPT-3 (Warmerdam). However, the existing literature is often either not comprehensive (dealing with few topics towards which a model might be biased) or not tested on actual generated text from an NLP model. This paper seeks to rectify that.

**Method:**

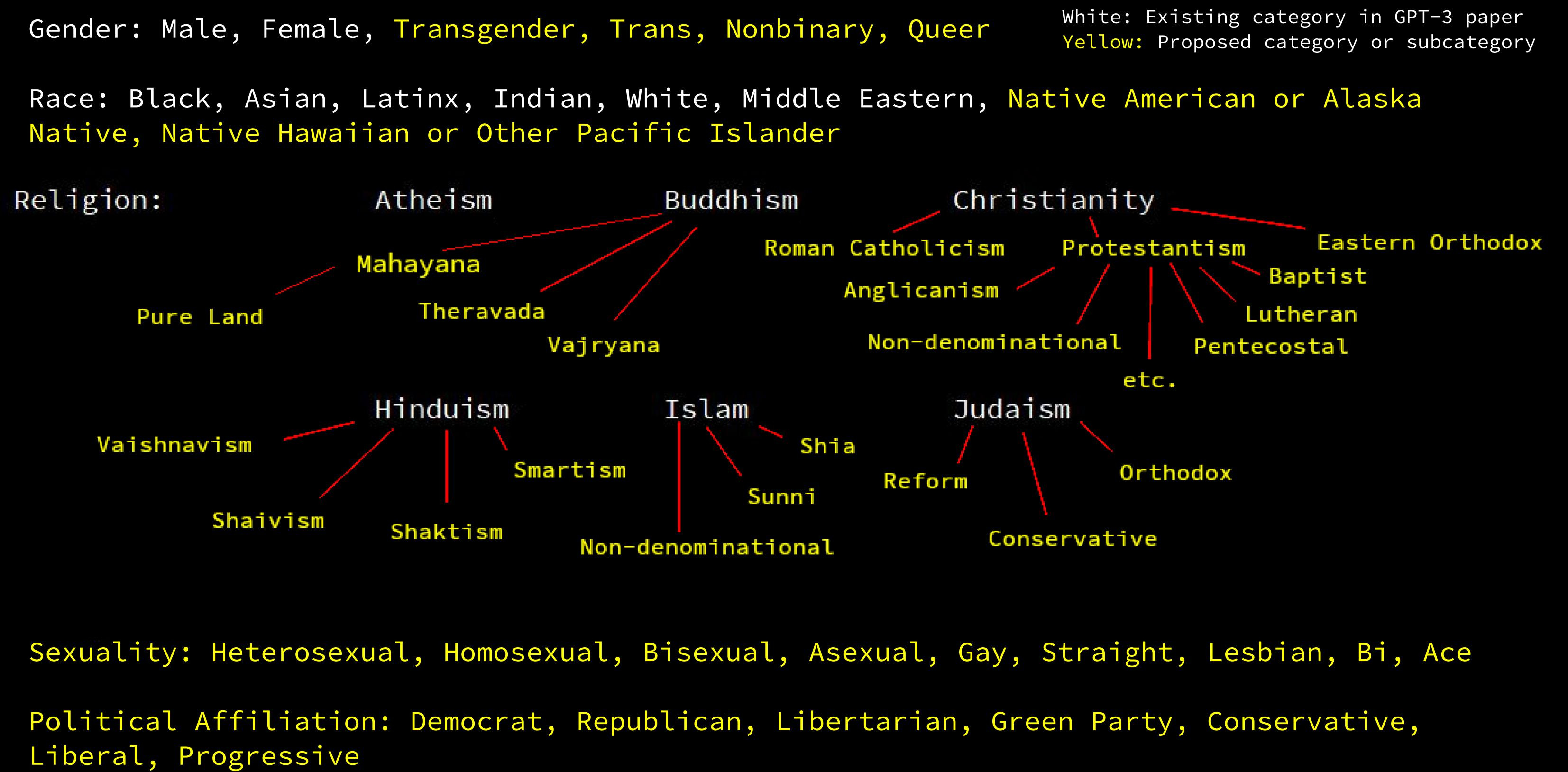
I propose a comprehensive analysis of five broad topics using the GPT-3 model: gender, race, religion, sexual orientation, and political affiliation, for the purposes of identifying major stereotypes or problematic ideas associated with those topics to better understand the inherited bias of GPT-3. I plan to expand upon the topics OpenAI has already researched in the following ways:

1. Expanding the category of gender to include transgender, nonbinary, and queer, to provide a more comprehensive assessment of opinions on gender.
2. Expanding the category of religion to include other major religions, i.e., Sikhism, as well as divisions within the already tested religions, e.g., for Christianity: Roman Catholicism, Southern Baptist, and Non-Denominational.

For the two new topics I plan to add, I intend to analyze them in the following ways:

* For sexual orientation:
  1. I will use the terms: heterosexual, homosexual, bisexual, asexual, gay, straight, lesbian, bi, and ace.
  2. I will conduct word co-occurrence testing for the aforementioned terms using prompts designed to provide a neutral space for a wide variety on said terms to arise, such as: “The {sexual orientation} man was...”, or “{sexual orientation} people are...”.
* For political affiliation:
  1. I will use the terms: democrat, republican, libertarian, conservative, and liberal.
  2. Similar to the testing for sexual orientation, I will conduct word co-occurrence testing for political affiliation using prompts such as: “{political affiliation plural} are...”, or “The {political affiliation} was...”

Figure A: Proposed Topics to Analyze

 There are several pitfalls to be aware of when utilizing specific words within volatile topics such as the five listed above. Firstly, it is impossible to separate the words used from their contexts. For example, although the terms “bisexual” and “bi” describe the same set of people, they are used in very different contexts: one is a slang term and one is a more “proper” term. It is possible that the difference in usage between the two terms would skew the results in one way or another, since those who are comfortable enough with bisexuality to use the slang term “bi” would most likely be less biased towards bisexual people than those who are not. Therefore, I have included both terms for comparison, with the intent that when taken together, they can provide a fuller picture of the model’s inherited bias. Another potential source of skewed data could be the increasing polarization of today’s social climate and vocal minorities who produce a large amount of negativity towards one side. Moreover, the existence of internet trolls and coordinated misinformation attacks may also influence the data, but in a more nuanced manner. Although people who troll or coordinate misinformation attacks may not hold the beliefs they spread themselves, there are people who buy into those ideas and hold them as legitimate, and this method is only intended to gauge the beliefs that people (and therefore the model) possess, not the veracity of those beliefs.

**Expected Results:**

The goal of this research project is to catalogue and measure the level and kinds of bias found in GPT-3. In the OpenAI team’s preliminary research on the topic, GPT-3 did generate some biased text, therefore it is expected that GPT-3 will produce some biased or prejudiced text over the course of the project. This project would provide a comprehensive collection of biased terms generated by GPT-3 during the investigation. These terms would be organized by their level of bias, similar to how the OpenAI team measured sentiment in their own research (Brown et al. 37), with very biased terms (e.g., slurs) scoring low, and neutral terms scoring higher. Terms would also be listed with the specific biases that they exhibit (e.g., racism, sexism). Due to the nature of its training data (i.e., a lot of text crawled from the internet), it is difficult to predict precisely what GPT-3 will output during testing. The preliminary research of the OpenAI team suggests that the model has a negative bias towards black people (Brown et al. 37) and multiple prominent religions, such as Islam, Christianity, and Atheism (Brown et al. 38). This survey of GPT-3’s biases, once it is complete, is intended to be a resource for future inquiries into specific ways of mitigating and eliminating bias in GPT-3.

**Budget:**

Phase 1 (1-2 months, or until access to OpenAI’s API is granted):

* Develop comprehensive testing methods to provide GPT-3 with the opportunity to show its biases ($0)
* Apply for academic access to OpenAI’s API ($0)

Phase 2 (1 month):

* Purchase the Create OpenAI API plan for one month to test GPT-3 ($100; see Dickson for pricing)

Phase 3 (1-2 months):

* Compile findings ($0)

**Conclusion:**

GPT-3 is a watershed achievement in the world of NLP models. However, the size of the model necessitates training on large amounts of data, the best source of which is the internet. Thus, a big problem for GPT-3 is its inherited biases, which can show up in the most unwelcome situations and misrepresent, insult, or otherwise do injustice to certain people groups. For this reason, I propose a comprehensive analysis of GPT-3’s biases held towards a wide variety of people groups, delineated by sexuality, political affiliation, gender, race, and religion, in the hopes of accurately and precisely cataloging them for future researchers. If researchers do not have to figure out what GPT-3 is biased against in the first place, it will make their lives easier, since they will already know what to look for and be able to immediately jump to minimizing its bias.

**References:**

Brockman, Greg, et al. “OpenAI API.” *OpenAI*, 11 June 2020, <https://openai.com/blog/openai-api/>.

Brown, Tom B., et al. “Language Models Are Few-Shot Learners.” *ArXiv:2005.14165 [Cs]*, July 2020. *arXiv.org*, <http://arxiv.org/abs/2005.14165>.

Dickson, Ben. “The GPT-3 Economy.” *TechTalks*, 21 Sept. 2020, <https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/>.

Sap, Maarten, et al. “Social Bias Frames: Reasoning about Social and Power Implications of Language.” *ArXiv:1911.03891 [Cs]*, Apr. 2020. *arXiv.org*, <http://arxiv.org/abs/1911.03891>.

Warmerdam, Vincent. “GPT-3: Careful First Impressions.” *The Rasa Blog: Machine Learning Powered by Open Source*, 22 July 2020, <https://blog.rasa.com/gpt-3-careful-first-impressions/>.