Worksheet: Identifying harmful bias in code, data collection, and algorithm development[1]

Purpose: This worksheet is intended to help AI researchers and developers seeking to practice the critical thinking necessary to avoid creating or reinforcing harmful and unfair[2] bias when developing AI applications. Please use this worksheet after reading through the key practices from our **Guide for Responsible Language in Artificial Intelligence & Machine Learning**.

Instructions: Complete activities A, B and C. Answers and key takeaways are provided at the end.

ACTIVITY A: Spotting harmful or unfair language in code

Certain terms used to document code have unfair or harmful origins, histories and/or connotations.[3] Review the blocks of code below and answer the following questions.

A.

```
def openpty():
    """openpty() -> (master_fd, slave_fd)
    Open a pty master/slave pair, using os.openpty() if possible."""

    try:
        return os.openpty()
    except (AttributeError, OSError):
        pass
    master_fd, slave_name = _open_terminal()
    slave_fd = slave_open(slave_name)
    return master_fd, slave_fd
```

B.

```
service cloud.firestore {
  match /databases/{database}/documents {
    function isBlackListed() {
      return exists(/databases/$(database)/blacklist/$(request.auth.uid))
    }

    match /posts/{postId} {
      // allow writes if user was never blacklisted or blacklisted more than 2 days ago
      allow write: if !isBlackListed();
      allow write: if request.time > get(/databases/$(database)/documents/blacklist/
        $(request.auth.uid)).data.blacklisted_at.value + duration.value(2,'d');
      allow read: if true;
    }

    match /blacklist/{entry} {
      allow read: if false;
      allow write: if false;
    }
```

1. Can you identify <u>at least one</u> term in each block of code that has a harmful origin, history and/or connotation? Circle the ones you identify.

2. Are these terms important to identify and replace? Why?

3.For each term you identified, what could be an alternative or replacement option? [Note: There is no one "right" answer!]

ACTIVITY B: Applying critical thinking about data and language

The following scenario is **fictional and features a fictional organization, the Health Data Institute**, but is based on true events. Read the scenario (in orange) and answer the questions related to it:

In March 2020, the World Health Organization declared COVID-19 a pandemic. Recognizing the importance of collecting data on the spread of the virus, a United States organization — the Health Data Institute — began tracking COVID-19 data across the country. Early on, a team within the Health Data Institute recognized disparities in COVID-19 health outcomes faced by different demographic groups. To better track and understand these disparities, the team decided to disaggregate by race the incidence and death rate data being collected. However, problems ensued shortly after. The team had included an array of labels for various racial groups that individuals were able to choose from in self-reported data collection forms, but the labels offered on the form resulted in confusion and questions. For instance: individuals who identified as Black Dominicans had to choose between "Black" and "Hispanic" on the form, but couldn't select both.

The team also worked with healthcare organizations in some states to consolidate data that was already being collected — but this meant relying on pre-existing forms and label options which varied, resulting in inconsistencies. For example, the Native Hawaiians and Pacific Islanders (NHPI) racial group was included on some forms as part of the "Asian" category, on some forms within an umbrella category of "other", and on some forms it was not listed at all.

1. What are the potential issues with how data related to race was collected and categorized by the Health Data Institute?

2. What are the potential healthcare impacts (including physical and mental wellbeing) of categorizing individuals using the labels listed in this scenario?

3. Can you think of other examples in which inconsistent or ambiguous labels have been used to capture people's identities? What are the impacts?

ACTIVITY C: Applying critical thinking about algorithms and language

Follow along with the scenario (in orange) and answer the questions related to it:

You are working at a startup that partners with other organizations to build custom NLP solutions for their product ideas.

Your team is currently choosing between three different product pitches. Before deciding on which tool to build, your team wants to think about potential ethical implications of building each of them. Here is the information you have about each product:

- **Tool A: A sentiment analysis model[4] that assesses the grammar and tone of blog pieces and provides suggestions for improvement. Users would install a browser extension and opt to use this tool to do a grammar and tone check prior to publishing their work.**
- **Tool B: A hate speech detection tool for a popular social media site. This tool would flag social media posts that contain hateful and/or offensive content, as understood by the algorithm your team builds.**
- **Tool C: A chatbot for a popular social media site. This tool would post on the site and attempt to converse with its human users, not just responding to messages but also telling jokes and stories. It would need to learn and get 'smarter' as it interacts with more people.**

1. For each of the above NLP tools, what might be potential issues related to racial bias through language that could come into play? Write potential issues for each below.

    a. Tool A:

    b. Tool B:

    c. Tool C:

Language models which analyze patterns of human language and power NLP systems, learn from huge amounts of digitized information. Regardless of which tool your team decides to build, the plan is to use the GPT-2 language model and dataset to train it. This dataset draws from vast amounts of digitized information including books, Wikipedia pages, blogs, news articles and more.

2. What may be some limitations of this dataset? Given the sources of information in the dataset, how might the dataset contain human bias, prejudice or toxicity?

3. What are some ways you can mitigate this bias, prejudice or toxicity in the dataset?

Your team builds and deploys Tool A (a sentiment analysis model that assesses the grammar and tone of blog pieces and provides suggestions for improvement). The tool works very well and is immediately picked up for use by individuals across the United States. However, you soon receive the following complaint from Jarvis, a Black food blogger.

Jarvis was writing a Twitter thread on the connection between Black history, identity, and food culture. Before publishing his Tweets, he decided to run his content through your AI tool. Here's what he drafted:

"When throwing down in the kitchen, it ain't enough to just know the recipe. Black food is called "soul food" for a reason. You gotta feel the connection between the ingredients, our ancestors, and our history. Mac and cheese, greens, yams, jerk chicken, and pound cake — these ain't just foods. They are central to us."

Your platform flagged Jarvis' writing, saying his language is incorrect, and his tone is angry.

4. Why might the algorithm have classified Jarvis' Tweets as being angry?

5. What might the algorithm be flagging as incorrect and why?

6. How might your team update the tool to mitigate the issues that this complaint surfaces?

Answer Key

Activity A

1. The problematic terms in each block of code / documentation are:

a. Master, slave
b. Blacklist / blacklisted

2. The terms identified here fall into one of the two categories we include when we talk about terms that have harmful origins, histories, and/or connotations: (1) terms with origins that are discriminatory or derogatory for certain racial groups / ethnicities; and (2) terms that insinuate or connect blackness with bad versus whiteness with good.

While the terms mentioned in this worksheet may not all be used in purposefully derogatory or discriminatory ways today, they don't exist in a vacuum. Terms such as "master/slave" can evoke painful, offensive meanings, particularly for Black people, negatively impacting their psychological well-being.[5] Terms such as "blacklist" conflate blackness with bad. Given the mutually reinforcing links between language and reality, such terms also perpetuate harmful, racist stereotypes and ideologies in society.[6] (**See this tab of our Terminology Guide for more examples and context**)

3. Companies including Google,[7] Apple,[8] Drupal,[9] and Linux[10] have started removing harmful terms from their coding platforms, shifting to using more inclusive language in their code and documentation.

Some alternatives for the problematic terms in this activity may include:
a. Master, slave
b. Blacklist / blacklisted

Activity B

1. The labels used here are inconsistent. "Hispanic", for instance, is listed as an ethnicity in other government forms — when asked by the CDC to self-identify as either Black or Hispanic, Black Dominican individuals were not given an opportunity to capture the full spectrum of their identities. Further, non-specific or umbrella categories like "other" are imprecise, conflating the experience of multiple racial groups that actually face significantly different health outcomes.

   It is important to acknowledge that the task of categorizing humans on the basis of demographic criteria such as race and ethnicity is complex. There are multiple conceptualizations of race and ethnicity — some view the two as distinct yet overlapping terms, while others view race as a subset of ethnicity. See our **Guide** for more information. There may not be a universal, prescriptive solution to this problem, but it is imperative to understand the real world impacts of generating and using race/ethnicity labels, and to be as precise as possible.

2. This COVID-19 example was problematic from a healthcare standpoint because Black Hispanic people (which include Black Dominicans) have been known to have different outcomes from White Hispanic people. So being labeled only under "Hispanic" would be imprecise and make it difficult to adequately track and account for the COVID-19 disparities between Black Hispanic and White Hispanic people.

3. **Incorrect categories can mask important social disparities between racial groups.** For example: different government databases use at least seven different terms to identify / label Native Americans, including terms like Native Americans, American Indian/Alaska Native, among others. Federal and state statistics also tend to misclassify Native Americans under other race / ethnicity categories. This has resulted in an undercounting of the Native American population and inaccurate reporting of key indicators that are used to allocate federal resources.[11]

   Using vague, **ambiguous terms can diminish the experiences of specific racial groups**. For example: in the 2020 election exit polls, CNN used the labels "White", "Black", "Asian", and "something else" to record voted information.[12] "Something else" conflates the experience of multiple racial groups, and fails to credit the powerful impact that specific voter demographics (Indigenous groups in particular) had during the election.

   Incomplete or insufficient categories can also cause erasure of already marginalized communities. Gender is a prominent example — most datasets only try to capture male and female gender labels, failing to incorporate the rest of the gender spectrum. This compounds the erasure of non-binary and transgender individuals, impacting their physical and mental health and wellbeing.

[NOTE] The scenario described in this activity is based on true events. The following articles contain additional details about the issues with COVID-19 data collection and labeling explored here:

- Meraji, S. M. (2020, April 22). The News Beyond The COVID Numbers. Retrieved from https://www.npr.org/2020/04/21/840609912/the-news-beyond-the-covid-numbers
- Ramirez, R. (2020, December 14). How Pacific Islanders have been left to fend for themselves in the pandemic. Retrieved from https://www.vox.com/2020/12/14/22168249/pacific-islanders-native-hawaiians-covid-19-pandemic

Activity C

1. The proposed tools could embed racial bias through language in the following ways:

    a. **Sentiment analysis used in NLP can result in computers picking up implicit, biased associations between race and particular words or categories.** For example, White-sounding names have been shown to be more deeply embedded with positive words like "love" and "laughter" than African American sounding names. Overall, African American names were associated with unpleasantness whereas European American names were associated with pleasantness.[13] Bias against Black-sounding names in the sentiment analysis tool could result in any content featuring Black-sounding names being associated with a more negative connotation, which could lead to disproportionate censorship of content featuring Black individuals.

    b. **Toxicity and hate speech are nuanced and hard to detect, making labeling hate speech a subjective process that can allow for bias.** What is considered hate speech evolves and depends on contexts. So, slurs that have been reclaimed by groups against whom they were originally weaponized can incorrectly be classified as hate speech. Hate speech algorithms also more often incorrectly identify African American English as hate speech (hateful or offensive) than "Standard" American English. This can lead to disproportionate censorship of African American writers.

    c. **Chatbot models use dynamic learning algorithms, which means that in addition to picking up on explicit and implicit biased associations within human language, they also give developers less ongoing control over the degree of bias in the language they learn from over time.** For instance: in 2016, Microsoft released a chatbot called Tay on Twitter, which was meant to experiment with "conversational understanding". While it was trained on "modeled, cleaned, and filtered" public data during its development phase, once it was deployed, Tay continued to learn from the interactions it was having with users. In less than 24 hours, the chatbot was producing racist, transphobic, and other discriminatory ideologies it had picked up from the humans it was observing.[14]

    d. OpenAI's GPT-2 training data includes vast amounts of Internet data including controversial articles posted to Reddit, which were later banned by Reddit for violating hate speech rules. Given that patterns of human prejudices and biases in these text sources become embedded in the NLP tools that learn from them, GPT-2 is prone to generating racist, sexist or otherwise toxic language.[15] Many AI language systems are built with as much data as possible, under the idea that more data makes the system more powerful and accurate. However, large language models run high risk of embedding bias and if datasets include abusive language, there can be significant negative impact. To read more about bias in language models, read our **Guide for Responsible Language in AI / ML**.

2. It is important to recognize that it is not possible to make these datasets and AI systems completely free of bias. When using a dataset such as GPT-2, carefully examine the quality of the data and the potential bias and toxicity. Recognize the sources of the data, and how those texts may contain biases. Then, document this information using tools like **data statements for NLP**. Be transparent about limitations.

If utilizing such existing datasets or language modes, prioritize and insist on documentation as part of dataset creation.

3. The word "jerk" generally has a negative connotation or is considered offensive. It may have been labeled or flagged as offensive. The algorithm picked up on that word without picking up on the context. It did not know that jerk chicken is a food dish and classified the whole essay as sounding negative.

4. The algorithm flagged the term "ain't" as incorrect. However, Jarvis is using African American English in his post and words like "ain't" and "double negatives" are part of the systematic grammar of African American English. There is nothing linguistically incorrect or inferior about these features. However, language datasets tend to contain an overrepresentation of "Standard" language varieties (like "Standard" American English). Other language varieties present in these datasets tend to be erroneously labeled as incorrect, or "unintelligible" — especially if annotators don't speak these varieties. As a result, African American English is often misclassified by AI systems.

5. Action can be taken related to the selection of data for the AI system to be trained on, the documentation of datasets, as well as the labeling of data.

    a. *Selecting data and datasets:*
        - Work to use balanced and representative training datasets that both incorporate language varieties of target users and reflect the communities and contexts in which the AI systems are used.
        - Reflect on the preferences the AI system is learning and who it may reward. For example, are preferences for linguistic traits or communication behaviors based on a hidden preference for White, middle-class norms? Is "suitability" or "eligibility" of an individual being linked to their proximity to "Standard" American English?

    b. *Documenting datasets:*
        - Recognize what data the AI system is trained on, the sources of the data, and how those texts may contain biases and prioritize certain voices. Then, document this information (using tools like data statements for NLP) alongside the goals for developing the dataset and be transparent about the limitations.

    c. *Labeling data:*
        - Ensure data labelers speak the specific language varieties they are listening to, and are trained in how to counteract implicit biases. Training individuals who speak less commonly spoken language varieties (including minoritized languages, as well as regionally and socially specific language varieties that may differ from the "standard" variety) may be a more time intensive and costly undertaking. However, it is the most effective way to ensure that language data is accurately labeled.
        - In regards to labeling hate speech, build in more objective guidance for labeling hate speech data.

Endnotes

1 This worksheet is part of the Responsible Language Guide for AI & ML project. It was developed by a team of researchers from the Center for Equity, Gender, and Leadership (EGAL) at the University of California Berkeley, Haas School of Business. It benefited from invaluable feedback and contributions from practitioners at leading tech companies and in academia. The project has a particular focus on race and the United States, although it includes examples across identities and geographies. We respectfully acknowledge that this work has been developed at UC Berkeley, which sits on unceded Ohlone land.

2 Fairness can have various definitions. In referencing "unfair" bias in AI systems, we mean bias that results in unjust impacts on people. When it comes to AI systems, justice considers how certain groups are oppressed or marginalized in the particular context and explores how the AI system can advance equity, rather than perpetuate a status quo that may oppress or marginalize certain groups.

3 When we say, "terms that have harmful origins, histories and/or connotations", we are referring to terms with origins that are discriminatory or derogatory for certain racial groups / ethnicities; terms derived from discriminatory and racist legislation; and terms that insinuate or connect blackness with bad versus whiteness with good.

4 Sentiment analysis uses text classification to predict opinions, how an individual or group of people feels in a particular context and/or how they might react. It's often used in marketing and customer service. It can also be used in ML algorithms to analyze and tag messages in social media and social networks

5 Gee, Gilbert C., Ro, Annie, Shariff-Marco, Salma, and Chae, David. (2009). Racial Discrimination and Health Among Asian Americans: Evidence, Assessment, and Directions for future research. Epidemiologic Reviews 31: 130-151. DOI: 10.1093/epirev/mxp009.

6 Houghton, F., & Houghton, S. (2018). "Blacklists" and "whitelists": A salutary warning concerning the prevalence of racist language in discussions of predatory publishing. Journal of the Medical Library Association, 106(4). doi:10.5195/jmla.2018.490

7 "Google Chrome and Android Move Away from 'Blacklist' - 9to5Google." Google, Google, 9to5google.com/2020/06/12/google-android-chrome-blacklist-blocklist-more-inclusive/.

8 Apple Style Guide. (n.d.). Retrieved from https://help.apple.com/applestyleguide/#/apdaf2bc3367

9 emove usage of "blacklist", "whitelist", use better terms instead. (2021, January 14). Retrieved from https://www.drupal.org/project/drupal/issues/2993575

10 (n.d.). Retrieved from https://git.kernel.org/pub/scm/linux/kernel/git/torvalds/linux.git/commit/?id=49decddd39e5f6132ccd7d9fdc3d7c470b0061bb

11 https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf#page=33

12 Arauz Peña, P. 2020, November 9. "Alaskans react to CNN poll labeling Native voters 'something else'". Alaska Public Media. https://www.alaskapublic.org/2020/11/09/alaskans-react-to-cnn-poll-labeling-native-voters-something-else/

13 Hutson, M. (2017). Even AI can acquire biases against race and gender. Science. https://science.sciencemag.org/content/356/6334/183/tab-pdf.

14 Vincent, J. (2016). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. The Verge. Retrieved from https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

15 Gehman, S., Gururangan, S., Sap, M., Choi, Y. & Smith, N. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. EMNLP.

This worksheet was developed by the Center for Equity, Gender & Leadership at UC Berkeley Haas School of Business. It is an accompanying resource to the guide, **Responsible Language in AI & ML**.