Toward More Responsible Labeling of ML Training Datasets

For Data Labelers

**Note: there are two options for this lesson plan, catering to slightly different audiences.*

Who this is for: Product managers or team leads, to use with data labelers who classify humans or label language data in AI or ML

Pre-reading: Both the instructor and participants should read the **Responsible Language Guide for Artificial Intelligence & Machine Learning** first.

Background: Individuals who label machine learning training datasets that include information on humans can take concrete steps to avoid creating or exacerbating unfair biases. When such proactive, responsible steps toward fairness are taken, the language used in labels can avoid contexts of racial bias. This lesson plan is designed for facilitators to demonstrate ways in which thoughtful, inclusive and fair label language applied to human data can proactively advance fairness and equity in ML systems.

Target participants:

- Employees / contractors at data labeling vendors to whom tech companies source out data labeling tasks. While the good practices can be adapted to data about humans based outside the US, the research findings and activities are most relevant for US-based data.

***Persona**: This individual has little (non-expert) understanding of how algorithms / AI systems use the data they label to make decisions. They may have basic knowledge of diversity, equity, and inclusion topics, but little to no knowledge of language to advance equity and inclusion. They receive annotation guidelines from project managers and follow the instructions provided to perform their labeling tasks, prioritizing speed and volume of annotations.*

- In-house employees of tech companies (temporary employees, vendors, and/or contractors) performing specialized data labeling tasks. While the good practices can be adapted to data about humans based outside the US, the research findings and activities are most relevant for US-based data

***Persona**: This individual has mid- to expert-level understanding of how algorithms / AI systems use the data they label to make decisions. They have undergone general diversity, equity, and inclusion training within the tech organization they work for, but are not fully equipped to ensure that the language they use (while coding and otherwise) advances equity and inclusion. The individual performs data labeling tasks for complex workflows, and so receives specialized, in-house training. They interact with the engineers*

providing them with annotation guidelines to some degree, and are able to provide feedback on the nature of their labeling tasks.

Goals: (1) Equip participants with an understanding of the ways in which human interventions and decisions can result in unfairly biased datasets, particularly as they pertain to the data labeling stage; (2) enable participants to identify opportunities to avoid areas of unfair bias in their own data labeling tasks, including unintentional unfairness; and (3) provide participants with responsible practices to follow to counteract unfair bias in data labels that could result in harms for downstream users.

Preparation:

- Familiarize yourself with this lesson plan.
    - We recommend the entire lesson as a general responsible data labeling training for labelers. However, if your audience is only working on image data, you may choose to skip activity B. If your audience is only working on language data (speech / text / sign), you may choose to skip activity A. When doing so, make sure to adjust the talking points and session time to account for how long each activity is slated to take.
- Reach out to registrants ahead of time to give them an idea of what to expect.
    - Communication note: Since the first activity works best if participants are not made aware of its objectives up front, we recommend that any email or other communication to participants be kept high-level. See Appendix i for a draft email to send to registered attendees.
- Set up the physical room (or virtual space) so participants are able to break out into smaller groups of 3-4 for the activities.
- Prepare the facilitation materials (see list below) and ensure participants have their participant packets before the training starts (see materials below).

Materials:

- For facilitators: Printed copy of this lesson plan, screen, projector, **presentation deck** to accompany lesson plan (also in Appendix ii)
- For participants: A notepad, pen, and the participant packet which includes a printed/soft copy of the **Responsible Practices handout** (also in Appendix iii) and a printed/soft copy of the **Personal action plan template** (also in Appendix iv).

Note:

- *Bullets vs. Numbers / Letters:* In this lesson plan, anything that is bulleted is a talking point for the facilitator. While anything that is numbered or denoted by letters is an instruction for the facilitator.

Time: 150 minutes

Introduction [15 MINUTES]

1. Welcome participants to the session, and ensure that they each have required materials ready. Go over the agenda slide.
    - Welcome! We are excited to welcome you to this workshop around labeling data responsibly -- particularly image and language data about humans. To get warmed up, we are going to do some introductions and then jump directly into an activity. I'll be introducing the purpose and roadmap for the lesson after the first activity -- the reason

for this will become clear!

- This is meant to be an interactive learning experience, so please make sure you have a notebook and pen or pencil.

2. Do a quick round of introductions. Have everyone say their name, pronouns, one thing they hope to learn about today, and one thing that's bringing them joy at the moment. Model this by going first.
3. Tell participants that you will now work together to co-create a community agreement. Explain:

   - A community agreement includes rules of conduct for the training. Community agreements engage participants in developing these ground rules so they have ownership of the rules and are responsible for them—individually and as a group.

4. Ask participants: What are some good rules or principles we should set as a group for this space? Have participants share verbally or via chat (if virtual) and add their suggestions to Slide 4.

   a. Be sure to share your screen so participants can view Slide 4. For in person settings, use a flipchart and write "Community Agreement" at the top.

   b. Add some suggestions to the community agreement yourself. Such as:

      - Engage in brave conversations. (Be thoughtful about what you say and be open to the fact that even comments that have good intentions can have negative impacts. Try to understand others' perspectives and be open to feedback.)
      - Take space, make space. (Allow everyone to have time to share and react. Be brave and speak up, but also step back and allow others time to contribute.)
      - Share your own experiences and honor others' experiences.
      - Practice active listening
      - Honor confidentiality.
      - (For virtual settings) Keep video on to the extent that is comfortable or possible for you, and remain focused (refrain from checking email / scrolling online).

5. Ask participants: Is there anything you would like to take off this list, or anything else to add to it?

6. To conclude, tell participants:

   - Please raise your hand (or give a thumbs up in a virtual setting) if you agree to this community agreement. If in a physical space, tape the community agreement to the wall in a place where people can see it.

Activity A [25 MINUTES]

7. Present slides 5 & 6 and explain the set up for Activity A using the talking points below. ***Do not*** let them know this activity is intended to introduce the topic of how language used to label images can embed racial / gender bias.

   - [Present slide 5] We will kick this off with an activity! To begin, please list out the numbers 1 through 10 on your sheet of paper.
   - [Move to slide 6] Here is a list of adjectives that can be applied to images of humans that need to be labelled as part of an ML training dataset. These adjectives include: Aggressive, Angry, Athletic, Bossy, Caring, Competent, Confident, Emotional, Enthusiastic, Leader, Neutral, Playful, Professional, Smiling.
   - For the first half of this activity, I will show you 10 images of people. These images were

collected with explicit consent. Each image will remain on the screen for 7 seconds. Within that time, you must choose ONE adjective to describe / label the adult in the image from this list that will remain on screen throughout the activity. You can reuse labels.

- Each picture will be numbered, so write down your label next to the corresponding number on your sheet.
- Before we get started, I will give you 20 seconds to familiarize yourself with this list of adjectives, but note that the list will remain on screen throughout.
- Any questions?

8. Once all participants have understood the activity, give them the 20 seconds to read through slide 6.

9. After 20 seconds are up, present slides 7-16. Start a timer as you present, making sure that each slide / image is up for exactly 7 seconds.

10. Move to slide 17 and tell participants the following:

- We are now going to take a look at the labels associated with certain groups of pictures. These images were chosen in deliberate sets -- for instance, there are three images of people wearing suits (business professional attire) and smiling. All images here are from stock imagery, and the people who appear have given their explicit consent for their images to be used for the purpose we are using them today in this exercise.

11. On slide 18, ask participants the following. Have 2-3 people share.

- Are there any recurring, subtle themes in the way people are depicted in the images?
- What are the labels you chose for each image?

12. Build off of the discussion to share the following points on how labels for image data can be subjective, and what causes these trends:

- Images of women -- even women in business attire -- are more often described according to physical, subjective attributes like "smile" and "beauty", while images of men are more likely to be described according to professional attributes like "leader" or "management".[1] [2] When these trends show up in data labels, it can result in search results of "leader" and "management" to skew towards images of men, perpetuating this gender bias and stereotype.
- This is seen in words used to describe people of different races as well. A study examined how Black and White quarterbacks were described by sports publications and found that Black quarterbacks were described in terms of their physicality and perceived lack of mental prowess, while White quarterbacks were described as lacking physicality but being more mentally adept at the game.[3] It revealed that Black and White quarterbacks were described in ways that reinforced stereotypes of White and Black men.
- AI / ML models can learn from such stereotypes contained in data labels. These stereotypes can then be perpetuated at scale in product experiences for many people who use products built from these models.

13. Repeat the same question for slides 19-21, having 2-3 people share:

- Are there any recurring, subtle themes in the way people are depicted in the images?
- What are the labels you chose for each image?

14. Build off of the discussion to share the following points:

- *For the set of images of individuals with 'neutral' expressions* [Slide 19]: Certain qualities of humans can be open to interpretation -- this includes emotions. The problem with labeling emotions is that this process is prone to bias. For instance, research finds that

Black women's facial expressions and behaviours are more likely to be considered angry than similar expressions / behaviors of White women.[4] This baseless "angry Black woman" stereotype reflects racism in society, and can result in bias that impacts how an annotator labels a Black woman's face.[5] If and when facial recognition software is trained on this data, it can advance these harmful biases and stereotypes. For example, AI systems have interpreted Black NBA players as having more negative emotional states than their White colleagues when they had comparable expressions.[6]

• *For the set of images of frowning individuals* [Slide 20]: This is another example of how stereotypes can come into play when labeling emotions. Frowning men are more often seen as 'angry' as compared to frowning women, who are more likely to be considered upset or generally 'emotional'.[7] In the US, the "angry Black man" stereotype leads to Black men being disproportionately and erroneously viewed as threatening or aggressive compared to White men displaying similar expressions, mannerisms, or behaviours.[8] If and when facial recognition algorithms are trained on data with labels that reflect this bias, they may disproportionately classify faces of Black men as angry or menacing, and as a result, perpetuate the erroneous stereotype.[9]

Can labeling emotions be an objective process?
Psychologists, anthropologists, and other researchers argue that emotions cannot be neatly divided into a "small set of basic categories like anger, disgust, fear, happiness, sadness, and surprise". They also find that these emotions cannot be directly mapped onto human facial expressions in a consistent, measurable, and uniform manner. In other words: assigning labels to nuanced human emotions is a subjective, complex task.

Researchers at the AI Now Institute and across various universities have raised concerns about this subjectivity. Despite this, efforts to label images with human emotions so algorithms can assess facial expressions are ongoing.

Although this activity uses labels that involve interpreting emotions, we do so only to highlight the biases and stereotypes rampant in popular, widespread image databases.

Sources:
Excavating AI. (n.d.). Retrieved from https://excavating.ai/
Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (2010). AI Now Report 2018. AI Now Institute. Retrieved from https://ainowinstitute.org/AI_Now_2018_Report.pdf

• *For the set of images of parental figures with their children* [Slide 21]: This ties in to gender roles and expectations in society. Mothers are often expected or assumed to be primary caregivers. So they may be more likely to be labeled as nurturing or caring. For fathers, caretaking is seen as more of a choice. They may be more likely to be labeled as playful than mothers in similar settings.

• Regardless of what you chose individually, we want to reiterate that research shows us that stereotypes can show up in data labels and result in discriminatory AI outcomes. This amplifies and reinforces systemic discrimination in our society. It is important to be aware of how our own biases can impact how we label images of people.

15. Conclude the activity with the following points:

• Gender and racial stereotypes are prevalent all around us. These stereotypes inform

biases in our minds. Often, we are not aware of the biases we have. This is called unconscious bias, which we will soon delve into. When we label images of people - especially when we do so quickly - our minds pull from our biases (whether conscious or unconscious) and impact how we label images of people.

- Take ImageNet, for example [Slide 22]-- one of the most widely used image datasets in machine learning. Many of these images are labeled subjectively. For example, a photograph of a woman smiling in a bikini is labeled with adjectives that have negative connotations, including "slattern" and"slovenly" (derogatory terms meaning a sexually promiscuous or dirty and untidy woman), and a young man drinking beer is categorized as an "alcoholic."[10]
- When data with labels that have harmful bias built in are used by AI systems, they can reinforce and exacerbate those stereotypes and biases.
- Another challenge linked to labeling data is labeling identity. For instance, it is not always possible to infer someone's gender identity from their physical appearance. With this in mind, Google recently removed gender labels from its facial recognition software, opting to use non-gendered labels such as "person" instead.[11]
    - Similarly, racial identity cannot necessarily be inferred from physical attributes like skin colour.
- All of these points bring us to the topic of our session!

Introduction to bias [10 MINUTES]

16. Introduce the core topic of the lesson. [Slide 23]
    - This lesson will delve further into how language used to label data about humans -- both image data and data containing text / speech -- can be biased, and the harms such labels can cause.
    - In particular, we will focus on racial bias in language and labels. But of course, as demonstrated just now, gender and other forms of bias come into play as well.
17. Ask participants: thinking about the outcomes of Activity A, what comes to mind when you hear the term bias? Have some participants share their answers.
18. Present Slide 24 and build off the answers to explain:
    - There are multiple definitions and types of bias. We're going to focus on one particular definition.
    - Unconscious biases are cognitive shortcuts: split-second judgments. Humans experience biases all the time, often unintentionally and unconsciously. Our brains are wired to be biased, and these biases show up in the initial reactions we have.
    - Humans have 2 key modes of thinking: System 1, which is our automatic, quick instincts; and System 2, which involves more deliberate effort, agency, and choice. We need our System 1 thinking to organize and manage all the stimuli we constantly face, but this is also where our biases come into play.
    - We make our quick judgments based on personal experiences, education, upbringing and communities -- and the stereotypes and norms that accompany them.
    - Of course, we can have conscious biases as well. But when labeling data, particularly in time-sensitive situations, we tend to fall back on the quick System 1 thinking. We tend to label data based on what is familiar to us, and any stereotypes or biases we hold can come into play unchecked.
    - Sometimes, it's the words themselves that are problematic. Widely-used image

databases such as ImageNet have been found to contain labels that reflect personal biases of the labellers. These can include language that is experienced as derogatory, or use racial and gender-based slurs.

- In other cases, how words are used reflect biases: such as women being described in terms of their physical attributes, objectifying them while men are more likely to be labeled in professional terms, historical stereotypes that can lead to harmful product experiences for people.
- Now you know how bias can affect the language labellers use to label image data, let's try another activity to see how it can come into play when labeling language data. Remember, it is important to recognize that we all have biases on account of system 1 thinking, and we want to be able to address them.

Activity B [20 MINUTES]

19. Introduce activity B and present slide 25:
    - Our first activity focused on images of humans, but biases can come up in labeling other types of data too - including language data. We'll now take a look at how this happens. I will read out a scenario and then have you break out into smaller groups of 3-4 to answer questions pertaining to the case.
20. Present the fictional scenario on slide 26. Read it aloud to participants, or invite participants to read it to themselves.

Jarvis, a Black food blogger, plans to write a twitter thread on the connection between Black history, identity, and food culture. Before he publishes the tweets, he decides it would be good to run his content through an AI powered writing tool that can catch spelling and grammatical errors, and detect the writer's tone of voice. Jarvis writes the following into the AI powered grammar tool, "*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.*"

Jarvis is shocked when the platform says that the tone sounds angry, and flags some of his language as incorrect.

21. Have participants break out into smaller groups of 3 people and discuss the following questions for 5 minutes: Why might the algorithm have classified the tone of Jarvis' essay as angry? What might the algorithm be flagging as "incorrect" and why?
22. Bring the entire group back together and have 2-3 breakout groups share what they discussed.
23. Use the following points to explain what happened in this case.
    - There are two main issues here. First, the word "jerk" generally has a negative connotation or is considered offensive.[12] So, 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.[13]
        - It's possible that "pound" may have been flagged as violent, since it is a word sometimes used to describe physical fighting.
    - Second, the algorithm flagged the term "ain't" as incorrect. However, Jarvis is using

African American English in his post. African American English uses words like "ain't", as well as double negatives that we don't see in "Standard" American English.[14] African American English is a language variety (as is "Standard" American English").

• Importantly, all language varieties are linguistically equal. No language variety is linguistically better or more correct. Rather, each language variety follows its own sets of rules. Terms like "ain't" and double negatives are not incorrect linguistically.

• Despite this, "Standard" American English is often seen as the 'right' or 'proper' way of speaking. This is why we put "Standard" within quotations. It is not inherently standard. Rather, it has been granted this position.

• Language datasets tend to favor "Standard" language varieties (like "Standard" American English). Other language varieties present in these datasets (such as African American English) tend to be wrongly labeled as "incorrect," or "unintelligible". This happens especially if annotators don't speak these varieties. As a result, African American English is often misclassified as being "incorrect" by AI systems.[15]

• This creates situations like the one with Jarvis, where the system is unable to interpret the context to understand that Jarvis is purposefully using African American English to speak in a voice that resonates with his audience.

24. Share the following and ask participants:

• This example illustrates how bias in labeling language data comes into play. Algorithms rely on patterns in how we use and label language. But what we say and how we say it varies based on language variety, speaker, and context. If a certain language variety is labeled as incorrect or unintelligible, it can carry over to the algorithm's outputs and perpetuate bias against this language variety and people who speak it. If a word is labeled as offensive but the machine doesn't recognize it's not offensive in certain contexts, that can be problematic too.

• What are some other issues that might come up when labeling text and speech? What are the implications for AI systems?

25. Build off their answers with the following:

• Language identification.

- Issue: Because language datasets are often centered around "Standard" language varieties, African American English is (1) underrepresented in training datasets, and/or (2) erroneously labeled as incorrect or unintelligible.

- Implication: Texts written in African American English are routinely misidentified as being written in a language other than English, at a rate far higher than texts written in "Standard" American English.[16] This has various implications such as texts written in AAE not showing up as much in search algorithms.

• Hate speech detection.

- Issue: Labeling hate speech is subjective. What is considered hate speech evolves and depends on contexts.[17] For example, slurs that have been reclaimed by groups against whom they were originally weaponized can erroneously be classified as hate speech.

- Implication: Hate speech algorithms more often incorrectly identify African American English as hate speech (hateful or offensive) than "Standard" American English.[18] Black people aren't the only folks affected. For example, the LGBTQ+ community has also reclaimed terms that were previously used as slurs against members of their community. These nuances and contexts are harder for algorithms to pick up on.

26. Present slide 27 and summarize the ways in which bias can come into play in labeling data.

- Bias can come into play…
  - When labeling images of humans: as our first activity demonstrates, labels of humans can be subjective and therefore be subject to bias. Biases also come into play if labeling an individual's perceived gender or race.[19]
  - When labeling language / language varieties: as our second activity demonstrates, our biases can result in inaccurate interpretation or identification. This can stem from not being familiar with the language variety used or biases around what is the "correct" way of speaking.
- It's important to keep in mind that instructions you get also play an important role. If you are provided with imprecise or incomplete annotation instructions, this can open the door to bias in labeling images of humans and language data.

Break [10 MINUTES]

Responsible practices [30 MINUTES]

27. Tell participants we will now think about responsible practices in data labeling in order to proactively mitigate unfair bias and advance racial equity and inclusion in data labeling.

28. Have participants break up into groups of 3 to brainstorm and identify practices for labeling data responsibly. Give the groups 10 minutes to think about practices to mitigate the particular harms discussed and advance racial equity and inclusion through responsible data labeling more broadly.

29. Bring participants back into the larger group and have each group share. As recommendations are brought up, write them on a white board, flipchart, or slide 29 under the header "Responsible practices for labeling data."

30. Build off of their answers with the following responsible practices to implement in their data labeling tasks using slide 30.

- 1. Recognize that labeling images of humans or language data can come with subjectivity. In labeling images of humans, ask yourself questions such as: *Might these labels be interpreted as having negative or positive connotations? Are these labels gendered (e.g., are you labeling women with more appearance-related tags and men with more profession-related tags; are you labeling women in scrubs as nurses vs. men in scrubs as doctors)? In labeling language data, ask yourself: Am I labeling language varieties I am unfamiliar with as unintelligible?* These types of questions can help you examine how your own biases are coming into play.
- 2. Consult the data labeling guidelines provided to check if perceived racial or gender-based categorizations are necessary. If race or gender-based categorizations are not necessary, do not use them. It may be appropriate to use more general labels like "person".
- 3. If race and /or gender labels are required - such as for fairness testing or dataset balancing purposes…
  - First, reference self-identification of these social categories if possible. When dealing with people's identities, it is best not to make assumptions, so label them as they identify themselves.
  - When self-identification is unavailable, label features when possible (versus perceived race or gender). If your task includes perceiving gender or gender

presentation, label features such as facial hair, presence of makeup, clothing, skin tone, etc. For perceived race, label features such as skin tone.[20] Use standard scales (typically provided in the labeling guidelines) such as the Fitzpatrick Skin Type.

- If labeling race or gender is still required, state that you are using perceived rather than self-reported identity. Be transparent about criteria used to assign perceived gender to individuals. This applies to skin tone as well.

• 4. Be transparent about any assumptions you make outside of the guidelines provided. For instance, if instructed to label text samples as offensive or not without being given a clear definition of what "offensive" means in this context, along with a range of examples, outline the definition(s) of "offensive" from which you will draw your labels, as well as indicators / explanations for each definition.

- Documenting how the labels used are defined as well as any assumptions made outside of the guidelines provided can feed into datasheets / statements for datasets or data cards, and help standardize the process as well as identify potential biases early on.

• 5. When in doubt, ask. Where possible, follow up with project managers if annotation details or guidelines are unclear or allow for ambiguity. Ask for examples to clarify any doubts.

Wrap up [40 MINUTES]

31. Leave participants with these questions: Which of these responsible practices can / will you be taking with you to your upcoming projects? What are some questions you can raise when you're being given your annotation instructions / training? Present the Personal Action Plan template (see Appendix iii) as you go through these questions.

32. Direct participants to download their own copies of the Plan to fill in. Give participants 10 minutes to reflect on the questions and complete their individual Plans.

33. Have participants pair up in breakout rooms for 5 minutes and share their Personal Action Plans with each other.

34. Wrap up the session. Go around the room and have each participant share:

a. What's one learning from the session that surprised you?

b. What's one word that describes how you're feeling right now?

35. Thank participants for joining the session. Provide the link to the immediate post-training survey. If there is time remaining, have them fill it out before leaving.

Facilitator Notes

1. Conduct an immediate follow up survey (see Appendix v) to understand the effectiveness of the training, as well as support participants to implement the good practices and lessons.

2. Develop a structured check-in process to follow up with participants after this lesson. This could look like the following:

- After 1 month: Use a survey to collect the following information from participants:
  - What good data labeling practices have you been able to implement?
  - How did that go and what did you learn?
  - Do you have any questions remaining around labeling datasets and potential bias?
  - What other support would you like?
- After 6 months: Use a survey to collect the following information from participants:
  - How have you actively made use of language that advances racial equity and inclusion in your labeling tasks?
  - Are there any challenges you have faced?
  - What have you learned?

Appendix

For facilitators

- Appendix i. Communication notes: A standard email (to be sent out upon registration, or a week prior to the session) could go as follows (update as needed with logistical information):

  "Thank you for registering for this important lesson around responsible data labeling. In this 90 minute session, I will guide you through a set of activities to showcase how human decisions in the labeling process can lead to bias. We will also explore opportunities to use labels that advance equity and inclusion instead. You will leave with good practices to implement in your labeling tasks."

- Appendix ii. **Deck accompanying lesson plan**

For participant packet

- Appendix iii: **Responsible Practices handout**
- Appendix iv: **Personal action plan template**

Feedback survey templates

- Appendix v: **Immediate follow-up survey**

This case study is an accompanying resource to the guide, **Responsible Language in AI & ML**. It was authored by Ishita Rustagi, Genevieve Smith, and Julia Nee with the Center for Equity, Gender & Leadership (EGAL) at the UC Berkeley's Haas School of Business. It benefited from invaluable feedback and contributions from practitioners at leading tech companies. It also benefited from prototyping and valuable feedback from: Alan Cummins, Aishani Patwari, Dominique Wimmer, Mairead Gordon, Paul Nicholas, Siobhan Hanna, and Tommy Denby, as well as a selection of TELUS AI's global community.

Endnotes

1 Simonite, T. (n.d.). When AI Sees a Man, It Thinks 'Official.' A Woman? 'Smile'. Retrieved from https://www.wired.com/story/ai-sees-man-thinks-official-woman-smile/.

2 Scheuerman, M. K., Paul, J. M., & Brubaker, J. R. (2019). How Computers See Gender. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-33. doi:10.1145/3359246.

3 Mercurio, E., & Filak, V. F. (2010). Roughing the Passer: The Framing of Black and White Quarterbacks Prior to the NFL Draft. Howard Journal of Communications, 21(1), 56-71. doi:10.1080/10646170903501328

4 Walley-Jean, J. C. (2009). Debunking the Myth of the "Angry Black Woman": An Exploration of Anger in Young African American Women. Black Women, Gender + Families, 3 (2): 68-86.

5 Walley-Jean, J. C. (2009). Debunking the Myth of the "Angry Black Woman": An Exploration of Anger in Young African American Women. Black Women, Gender + Families, 3 (2): 68-86.

6 Rhue, L. (2018). Racial Influence on Automated Perceptions of Emotions. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3281765.

7 Shu-Tsen Kuo Author Profile. (2019, January 08). Gendered Emotions: Raging Men and Weeping Women: Lead Read Today. Retrieved from https://fisher.osu.edu/blogs/leadreadtoday/blog/gendered-emotions-raging-men-and-weeping-women.

8 Wingfield, A. (2007). The Modern Mammy and the Angry Black Man: African American Professionals' Experiences with Gendered Racism in the Workplace. Race, Gender & Class, 14(1/2), 196-212. Retrieved May 27, 2021, from http://www.jstor.org/stable/41675204.

9 Shapiro, J. R., Ackerman, J. M., Neuberg, S. L., Maner, J. K., Vaughn Becker, D., & Kenrick, D. T. (2009). Following in the wake of anger: when not discriminating is discriminating. Personality & social psychology bulletin, 35(10), 1356–1367. https://doi.org/10.1177/0146167209339627

10 Excavating AI. (n.d.). Retrieved from https://excavating.ai/.

11 Johnson, K. (2020, February 21). Google Cloud AI removes gender labels from Cloud Vision API to avoid bias. Retrieved from https://venturebeat.com/2020/02/20/google-cloud-ai-removes-gender-labels-from-cloud-vision-api-to-avoid-bias/.

12 Crabb, J. (2019, May 28). Classifying Hate Speech: An overview. Retrieved from https://towardsdatascience.com/classifying-hate-speech-an-overview-d307356b9eba.

13 This scenario is based on a real world situation in which an African American-Jewish culinary historian penned an essay on Black food history, resistance, and Black joy. When he ran it through Grammarly, the software's tone detector erroneously suggested that the tone of his piece was angry. He Tweeted about the incident, and representatives from Grammarly responded. They looked into the contents of his work and found that the software had misinterpreted the word "jerk" (as in "jerk chicken"). They issued an apology and claimed to have fixed the issue. Read more at https://twitter.com/Grammarly/status/1314348739822342146.

14 We call the variety of English that is commonly promoted in places like business, media, and education "Standard" American English. "Standard" American English is based on the language used by those in power and is not objectively better than any other language variety. For this reason we put "Standard" in quotes. This variety is also referred to as White dominant language. African American English (AAE) refers to the language varieties used by many Black descendants of enslaved people in the US. Black Americans speak a diversity of language varieties, but these varieties do share some traits, including how they are often unjustly treated by those in power.

15 Jurgens et al. (2017). Incorporating dialectal variability for socially equitable identification.

16 Blodgett, S. L. & O'Connor, B. (2017). Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English

17 Matsakis, L. (2018). To break a hate-speech detection algorithm, try 'love'. Wired. Retrieved from https://www.wired.com/story/break-hate-speech-algorithm-try-love/.

18 Davidson, T., Bhattacharya, D. (2020). Examining Racial Bias in an Online Abuse Corpus with Structural Topic Modeling. Retrieved from https://arxiv.org/abs/2005.13041.

19 Race cannot always be mapped onto or inferred from physical characteristics. So, perceived "skin tone" is often labeled instead. However, it is important to also be aware of the limitations / subjectivity of labeling skin tone. Read more here: Dixon, Angela R., and Edward E. Telles. "Skin Color and Colorism: Global Research, Concepts, and Measurement." Annual Review of Sociology, vol. 43, no. 1, 2017, pp. 405–424., doi:10.1146/annurev-soc-060116-053315.

20 While "skin tone" is often labeled instead of race, it is important to also be aware of the limitations / subjectivity of labeling skin tone. Read more here: Dixon, Angela R., and Edward E. Telles. "Skin Color and Colorism: Global Research, Concepts, and Measurement." Annual Review of Sociology, vol. 43, no. 1, 2017, pp. 405–424., doi:10.1146/annurev-soc-060116-053315.