Making Stakeholder Impacts Visible in the Evaluation Cycle: Towards Fairness-Integrated Shared Tasks and Evaluation Metrics

Emily M. Bender (@emilymbender)
University of Washington

CiML 2019: Machine Learning Competitions for All
NeurIPS 2019 workshop
13 December 2019

Acknowledgments

- This talk represents joint work with:
    - Hal Daumé III, University of Maryland
    - Bernease Herman, University of Washington
    - Brandeis Marshall, Spelman College

- This is a work-in-progress presentation, from the very beginning of a project

The safe and equitable deployment of AI requires bias mitigation

- If ML/AI technologies are to tackle grand challenges and benefit all segments of society, they must not be tools of oppression

- The social world is rife with unfairness and discrimination, so the development of AI systems must pursue *bias mitigation*

    - lest we perpetuate or exacerbate bias (Sweeney, 2013; Buolamwini and Gebru, 2018; Noble, 2018; inter alia)

    - just trying for “neutral” isn’t enough

- ***Bias:*** cases where systems “systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others.” (Friedman & Nissenbaum 1996)

Goals

- integrate considerations of how AI systems impact
    - both direct and indirect stakeholders
    - into the development and evaluation portions of the system development lifecycle;
- develop metrics and data sets to evaluate the impact of bias in technology as used in the real world;
- ultimately, generalize these concerns into a living “best practices” document that other researchers can use when developing new tasks.

Focus on natural language processing (NLP)

- Human language is intimately tied to identity
  - how we present ourselves in the world
- Human language is intimately tied to the construction of the social world
  - e.g. social roles maintained through language

Focus on natural language processing (NLP)

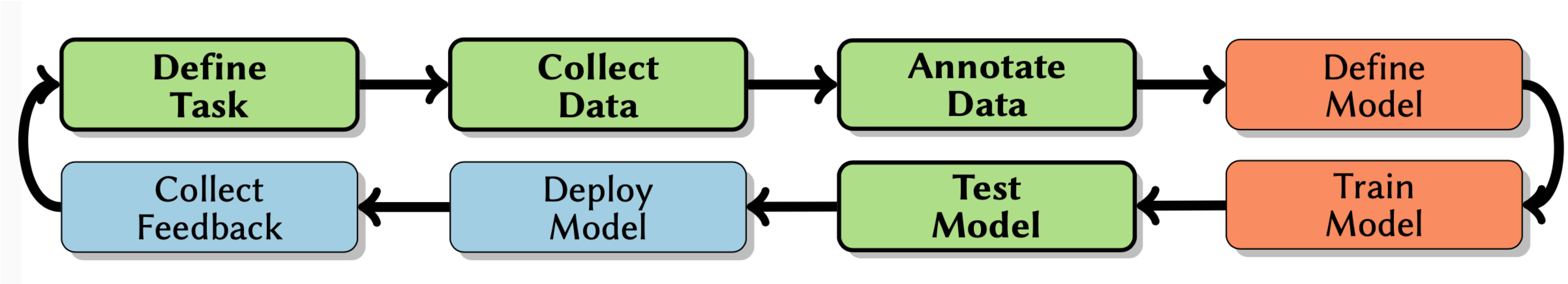
- Human language is intimately tied to identity
  - how we present ourselves in the world
- Human language is intimately tied to the construction of the social world
  - e.g. social roles maintained through language

=> Risks include some not found with other applications of ML

Goals

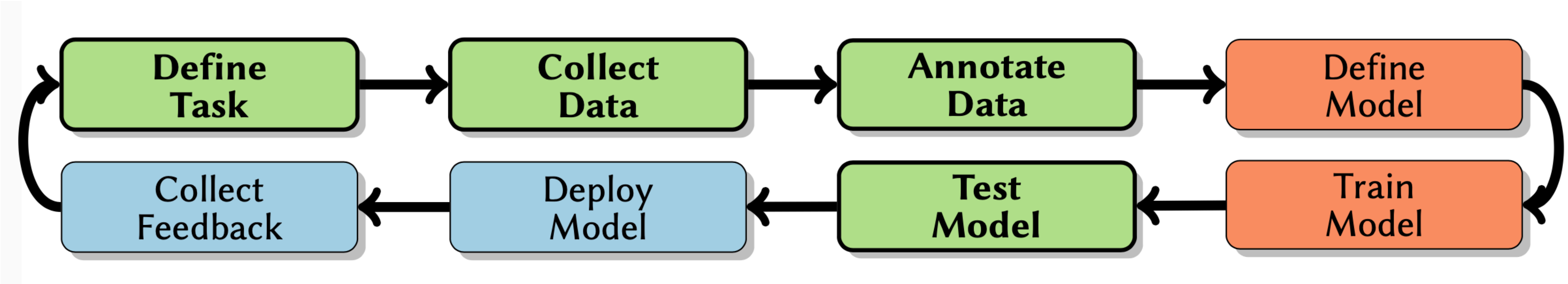
- integrate considerations of how AI systems impact
    - both direct and indirect stakeholders
    - into the *development and evaluation portions* of the system development lifecycle;
- develop metrics and data sets to evaluate the impact of bias in technology as used in the real world;
- ultimately, generalize these concerns into a living “best practices” document that other researchers can use when developing new tasks.

ML Life-Cycle (after Vaughan & Wallach 2019)



- As shared task organizers, we focus on:
  - Task definition
  - Data collection
  - Data annotation
  - Model testing

ML Life-Cycle (after Vaughan & Wallach 2019)



- As shared task organizers, we focus on:
  - Task definition
  - Data collection
  - Data annotation
  - Model testing

Biases can enter at any point in this lifecycle

Goals

- integrate considerations of how AI systems *impact*
    - both direct and indirect stakeholders
    - into the development and evaluation portions of the system development lifecycle;
- develop metrics and data sets to evaluate the impact of bias in technology as used in the real world;
- ultimately, generalize these concerns into a living “best practices” document that other researchers can use when developing new tasks.

One taxonomy of harms [Not mutually exclusive] (from Barocas et al, 2017; Crawford, 2017)

- allocational harms: ML systems unfairly allocating finite resources
- representational harms: ML systems contribute to subordination of certain groups
    - quality of service (e.g. ASR working better for some groups than others; Tatman, 2017)
    - stereotyping (e.g. online ads suggesting that people with Black-sounding names had been arrested; Sweeney, 2013)
    - denigration (e.g. Tay, where the ML system actively participated in hate speech; Price, 2016)
    - under-representation (e.g. image search for “CEO” returning more images of white men than is reflected in the real world; Kay et al, 2015)

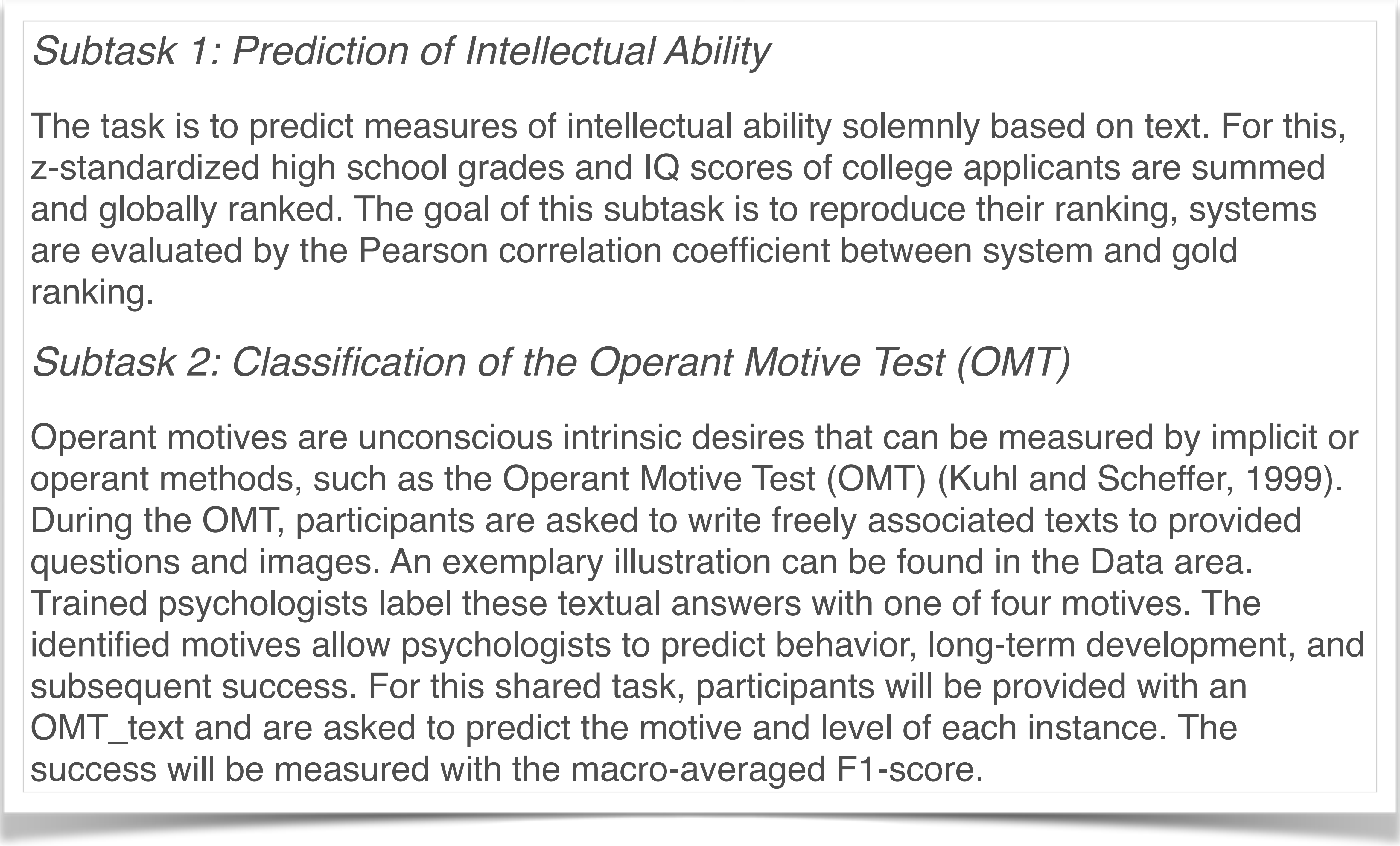
Goals

- integrate considerations of how AI systems impact
    - *both direct and indirect stakeholders*
    - into the development and evaluation portions of the system development lifecycle;
- develop metrics and data sets to evaluate the impact of bias in technology as used in the real world;
- ultimately, generalize these concerns into a living “best practices” document that other researchers can use when developing new tasks.

Responsibility of due diligence

- Shared task organizers direct the research efforts of larger groups
  - ==> an extra burden of responsibility to do ethical due diligence
- What are the use cases of the technology being developed?
- How does the specific ML task (inputs, outputs) relate to the intended use case?
- What are the failure modes and who might be harmed?
- What kinds of bias are likely to be included in the training data?

Instructive case study: GermEval 2020



Subtask 1: Prediction of Intellectual Ability

The task is to predict measures of intellectual ability solemnly based on text. For this, z-standardized high school grades and IQ scores of college applicants are summed and globally ranked. The goal of this subtask is to reproduce their ranking, systems are evaluated by the Pearson correlation coefficient between system and gold ranking.

Subtask 2: Classification of the Operant Motive Test (OMT)

Operant motives are unconscious intrinsic desires that can be measured by implicit or operant methods, such as the Operant Motive Test (OMT) (Kuhl and Scheffer, 1999). During the OMT, participants are asked to write freely associated texts to provided questions and images. An exemplary illustration can be found in the Data area. Trained psychologists label these textual answers with one of four motives. The identified motives allow psychologists to predict behavior, long-term development, and subsequent success. For this shared task, participants will be provided with an OMT_text and are asked to predict the motive and level of each instance. The success will be measured with the macro-averaged F1-score.

https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/germeval-2020-psychopred.html

Instructive case study: GermEval 2020

Instructive case study: GermEval 2020

- What could possibly go wrong?

Instructive case study: GermEval 2020

- What could possibly go wrong?

- (What could possibly go right??)

Instructive case study: GermEval 2020

> “In the United States, there is considerable evidence that IQ tests are racially biased. In the past, courts have excluded IQ tests from educational placement in California for precisely this reason. I wonder if there is research on this topic in the German context.
>
> “It is not difficult to imagine that the outcome of this shared task would be a set of technologies that encode spurious correlations between estimates of intelligence and the linguistic features of specific racial groups. If such a system were trained on data that already contains biases, there is a risk that this bias would be not only entrenched but amplified. And even if the IQ test statistics are not themselves biased, an NLP system that predicts IQ from text could introduce bias, if there is an unmeasured confound that is statistically associated with both IQ and race.”

(Jacob Eisenstein, message to corpora mailing list, 12/4/2019)

Questions that should have been asked

- Does the output of the ML task match the information it’s framed as predicting? (No.)
- Does the input to the ML task actually contain enough information to predict the output? (No.)
- What are the intended use cases for this technology?
- If the technology is working as intended, who might be harmed and how?
- If the technology is not working as intended, who might be harmed and how?

Questions that should have been asked

- Does the output of the ML task match the information it's framed as predicting? (No.)
- Does the input to the ML task actually contain enough information to predict the output? (No.)
- What are the intended use cases for this technology?
- If the technology is working as intended, who might be harmed and how?
- If the technology is not working as intended, who might be harmed and how?

Asking is the first step, but how to answer reliably?

Value Sensitive Design & Diverse Voices

- VSD (Friedman & Hendry, 2019): a framework for moving toward more beneficial and less harmful technology by explicitly identifying stakeholder values and then incorporating those values into system design
- Direct stakeholders: Those who will use or build the technology
- Indirect stakeholders: Others affected by the use of it

Diverse Voices (Young et al 2019) … applied to shared tasks

- Identify stakeholder groups
- Recruit panels of experiential experts
- Present the technology (here: shared task) in an accessible fashion
- Facilitate discussion with panelists to learn from their expertise about how the technology could affect them & their communities
- Analyze & summarize panel transcript and revise task definition accordingly — or abandon task all together
- Communicate revisions to panelists

Questions for panels

- How do you imagine using this technology?
- What benefits do you think it will have for you?
- How do you imagine others will use this technology?
- What concerns do you have about how the use of this technology might affect you/your community/society at large?
- What potential for intentional misuse do you see?
- What safeguards would you like to have in place?

Goals

- integrate considerations of how AI systems impact
  - both direct and indirect stakeholders
  - into the development and evaluation portions of the system development lifecycle;
- develop metrics and data sets to evaluate the impact of bias in technology as used in the real world;
- ultimately, generalize these concerns into a living “*best practices*” document that other researchers can use when developing new tasks.

Representative shared tasks: Desiderata

- Connect to real user or by-stander experience
- Potentially impact different marginalized populations
- Have a diversity of anticipated harms
- Have existing datasets and metrics
- Low barrier to entry for participants

Representative shared tasks: Desiderata

- Connect to real user or by-stander experience
- Potentially impact different marginalized populations
- Have a diversity of anticipated harms
- Have existing datasets and metrics
- Low barrier to entry for participants

Limitations:
Only focusing on English
Limited consideration of intersectionality

Goals

- integrate considerations of how AI systems impact
    - both direct and indirect stakeholders
    - into the development and evaluation portions of the system development lifecycle;
- develop metrics and *data sets to evaluate the impact of bias in technology as used in the real world*;
- ultimately, generalize these concerns into a living “best practices” document that other researchers can use when developing new tasks.

Proposed Task 1: Hate Speech Detection

- Toxic environments on social media can lead to the further marginalization of marginalized voices, and the affordances of social media sites make it easy for a small number of hate-promoting accounts to quickly create toxic environments (Daniels, 2017)

- Automating hate speech detection could help prevent this, but poor performance leads to *quality of service* harms

- Current SOTA systems perform far worse in production, due to:

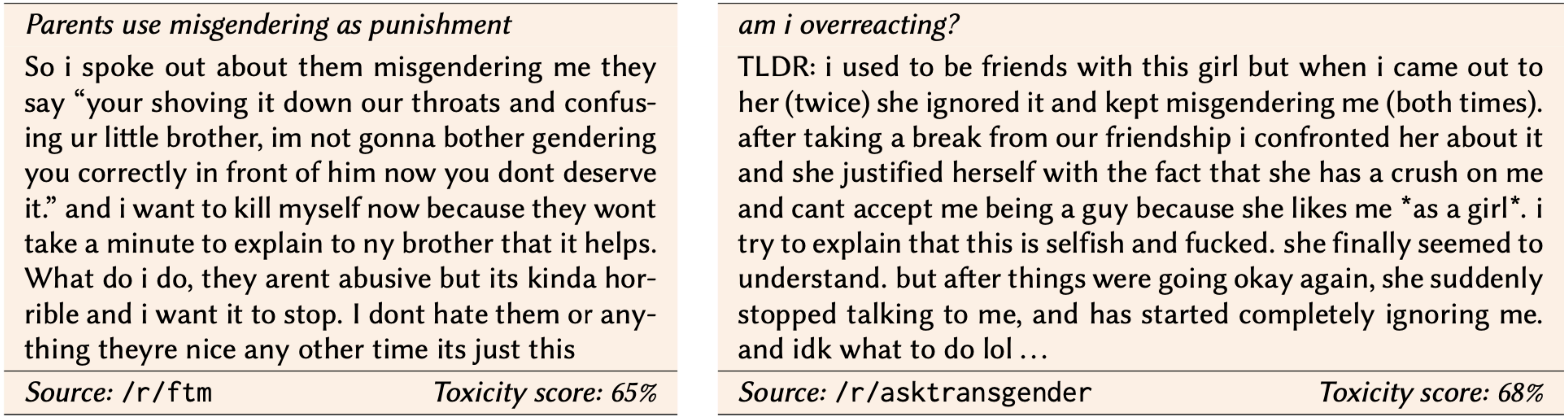
    - the way that shared task datasets are sampled (Wiegand et al, 2019)

    - biases included in annotation (Sap et al, 2019)

- Effect of false positives depends on whether the author of the flagged post is a member of a marginalized group (Dixon et al 2018; Sap et al, 2019)

Proposed Task 1: Hate Speech Detection

- Goal: Develop systems which help keep away hate speech without flagging posts where users describe an unpleasant event they experienced and ask for help



Parents use misgendering as punishment

So i spoke out about them misgendering me they say “your shoving it down our throats and confusing ur little brother, im not gonna bother gendering you correctly in front of him now you dont deserve it.” and i want to kill myself now because they wont take a minute to explain to ny brother that it helps. What do i do, they arent abusive but its kinda horrible and i want it to stop. I dont hate them or anything theyre nice any other time its just this

Source: `/r/ftm` *Toxicity score: 65%*

am i overreacting?

TLDR: i used to be friends with this girl but when i came out to her (twice) she ignored it and kept misgendering me (both times). after taking a break from our friendship i confronted her about it and she justified herself with the fact that she has a crush on me and cant accept me being a guy because she likes me *as a girl*. i try to explain that this is selfish and fucked. she finally seemed to understand. but after things were going okay again, she suddenly stopped talking to me, and has started completely ignoring me. and idk what to do lol ...

Source: `/r/asktransgender` *Toxicity score: 68%*

Table 1: Examples (from reddit) of users discussing distressing/harmful incidents that have affected their lives, both assigned more-likely-than-not probabilities of being toxic according to Google Perspectives toxicity classifier.

Proposed Task 1: Hate Speech Detection

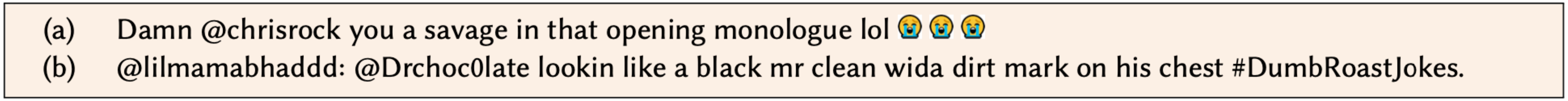
- Include both explicit and implicit abusive language
- Include items likely to be inaccurately flagged as hate speech and problematic if so flagged—individuals speaking about their own identity
- Proposed data source: Well-moderated subreddits such as /r/asktransgender, /r/gender, /r/ftm, /r/mtf/, /r/androgyny (seek permission from moderators)
- Training data: use moderator actions as noisy labels
- Dev/test data: further hand-annotation, incorporating the in-community insights of the moderators’ actions

Proposed Task 2:
Sentiment analysis for Black Twitter

- Sentiment analysis (Wiebe 1990, 1994, 2000):
  - classifying texts as expressing positive/negative/neutral sentiment,
  - in general or wrt specific aspects of some entity
- Component of systems that monitor public feedback on brands and public policy decisions (e.g. Chamlertwat et al., 2012; Desouza and Jacob, 2017)
- Black Twitter: the subcommunity of Twitter that uses specific hashtags and other means to produce a sense of community around racialized identity and social justice activism (Brock, 2012; Sharma, 2013)

Proposed Task 2:
Sentiment analysis for Black Twitter

- Off the shelf sentiment analysis systems perform very poorly on tweets from Black Twitter (Marshall, 2019)



(a)	Damn @chrisrock you a savage in that opening monologue lol 😭😭😭
(b)	@lilmamabhaddd: @Drchoc0late lookin like a black mr clean wida dirt mark on his chest #DumbRoastJokes.

Table 2: Black Twitter examples from: (a) 2016 Oscars (Marshall, 2019) and (b) #DumbRoastJokes (Florini, 2014).

- Likely causes of poor performance:
  - Language variation
  - Discourse practices such as *signifyin':* a practice which serves to perform Black identity in virtual spaces, and also involves layers of meaning that are opaque to outsiders (Florini, 2014)

Proposed Task 2: Sentiment analysis for Black Twitter

- *Quality of service harms*: Sentiment monitoring systems likely don’t capture the opinions of Black Twitter

- Sentiment analysis systems do perform better on Black Twitter would support work such as that of Bosley et al. (2019), which seeks to better understand the frequently erased leadership of Black women in social movements by analyzing Twitter discourse, mitigating *stereotyping harms*

Proposed Task 2:
Sentiment analysis for Black Twitter

- Recruit annotators from various subcommunities of Black Twitter to annotate by proposing their own tweets to include in each category of the annotation schema
    - Better information about original intent
    - Opt-in data collection
- Weakly supervised training set of tweets annotated with specific hashtags (e.g. #DumbRoastJokes) or emojis

Proposed Task 3: Visual Question Answering

- Input: Image plus textual question
- Output: Textual answer
- Open-ended and multi-modal
- Existing VQA systems show gender-stereotype bias (Hendricks et al, 2018) and tend to provide answers relating to objects not present but correlated to those that are (Rohrbach et al, 2018)
- Current image datasets are biased towards White American and European cultures (Shankar et al, 2017; DeVries et al, 2019; Ardalan and Malesky, 2018)

Proposed Task 3: Visual Question Answering

- VQA is being used as a basis for studying “commonsense reasoning” (e.g. Zellers et al, 2019)
- Geographically and culturally non-representative VQA datasets will lead to biased notions of “commonsense” and thus *denigration* harms
- Goal: A VQA task where
  - images come from diverse cultures
  - Q&A reflect the perspective of the culture the images come from

Summary

- These three tasks are designed to support evaluation of potential impact alongside accuracy
- Different types of harms addressed
- Different marginalized populations considered
- Different types of output (text classification, free-text generation)

Goals

- integrate considerations of how AI systems impact
  - both direct and indirect stakeholders
  - into the development and evaluation portions of the system development lifecycle;
- develop *metrics* and data sets *to evaluate the impact of bias in technology as used in the real world*;
- ultimately, generalize these concerns into a living “best practices” document that other researchers can use when developing new tasks.

Metric design: Desiderata

- Incorporate notions of fairness directly into the development cycle
- Center the impact on marginalized populations
- Extract best practices across metrics that handle different kinds of output formats

Metric design: Hate speech detection

- Harm-weighted metrics: The cost of an incorrect prediction should not only account for harm to the immediate affected party, but also the cost of further perpetuating harm toward that marginalized group
  - Possibly non-uniform costs of harm, depending on the speaker
- Inspired by preference-informed individual fairness (PIIF) (Kim et al., 2019)

Metric design: Sentiment analysis

- To address the complexities of sentiment analysis on Black Twitter, we must move beyond a simple positive/negative/neutral framing and take into account degree of sentiment, as well as sarcasm and *signifyin’*
    - More complicated label space ==> more complex metrics
- Possible factors to incorporate:
    - the dynamic range that systems are able to recognize
    - ability to handle quickly evolving vocabulary
    - ability to work across geographic regions and other subcommunities

Metric design: VQA

- Metric should account for different kinds of harms:
  - Errors related to cultural appropriation
  - Errors related to negative stereotypes
  - Errors from varying levels of specificity of the examples
- Further complexities arise when considering long, open-ended textual answers

Diverse Voices

- All of the above is initial ideas, to be refined through Diverse Voices panels
- Given participant values and concerns,
  - are the tasks appropriate to pursue?
  - are there more appropriate sources for train/dev/test data?
  - can the metrics be shaped to better measure possible harms?

Goals: Reprise

- integrate considerations of how AI systems impact
  - both direct and indirect stakeholders
  - into the development and evaluation portions of the system development lifecycle;
- develop metrics and data sets to evaluate the impact of bias in technology as used in the real world;
- ultimately, generalize these concerns into a living "best practices" document that other researchers can use when developing new tasks.