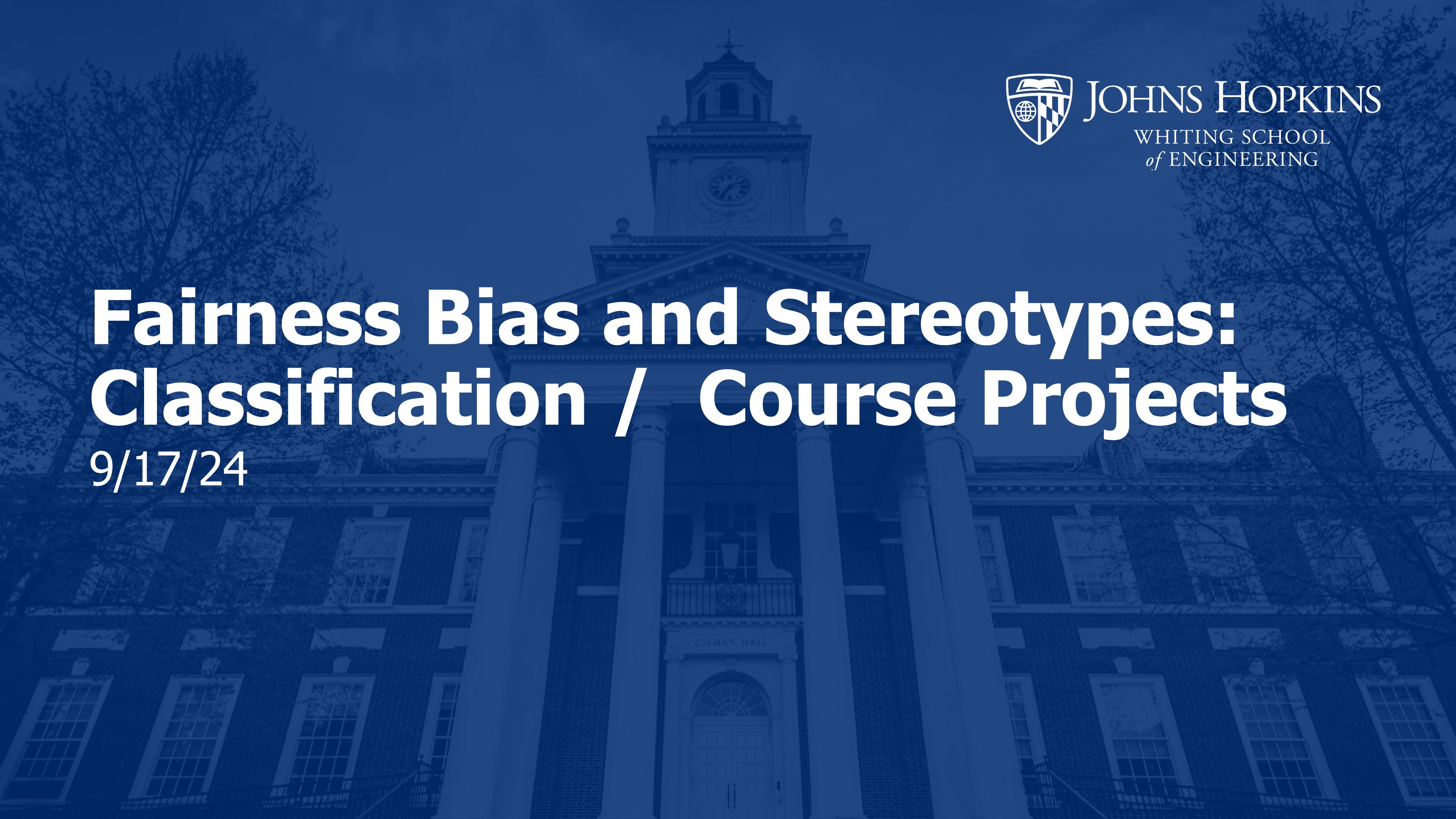


JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Fairness Bias and Stereotypes: Classification / Course Projects

9/17/24

Course Project

- http://ai-ethics-601-770.cs.jhu.edu/fa2023/project.html

1. Literature Review [Due September 27]
2. Project Proposal (4-8 pages)

   Written Report [Due October 25]

   Short in-class presentation
3. Final Report [Due at the end of the semester]

   In-class poster session

Literature Review

- Short paper (min 2 pages, max 6 pages) summarizing and synthesizing several papers (or works in general)
- Groups of one should review 5 papers, groups of two should review 7 papers, and groups of three should review 9 (you are welcome to review more)
- Most people will choose the same topic for your lit review and final paper, but this is not required, and the lit review will be graded on its own
- More tips on the course website: http://ai-ethics-601-770.cs.jhu.edu/fa2024/project.html

Please use the ACM template, with “nonacm” and “sigconf parameters (2-column format) https://www.acm.org/publications/proceedings-template. No need to include the teaser figure, CSS concepts, or keywords

Project Topics

- Can be anything related to the course
- We encourage you to talk to instructors about ideas (via email, Piazza, office hours, schedule additional appointment)
- Literature review focuses on identifying broad topic area – you have more time for specific project ideas!

Project Examples

- Conduct a technical assessment of a model or deployed system, such as probing for unfairness
  - Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. ACL 2023
  - Bianchi, Federico, et al. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale." FAccT 2023.
  - Inna Wanyin Lin*, Lucille Njoo*, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. "Gendered Mental Health Stigma in Masked Language Models" EMNLP (2022)

Project Examples

- Write a critical survey of a topic in AI
  - Language (Technology) is Power: A Critical Survey of “Bias” in NLP. Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Annual Meeting of the Association for Computational Linguistics (ACL)
  - Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. "A Survey of Race, Racism, and Anti-Racism in NLP" ACL (2021).

Project Examples

- Conduct a technical study, focused on building models to address ethical concerns.
  - Methods for reducing toxicity in model outputs
  - Methods for reducing stereotyping in model outputs
  - Methods for privacy preservation and anonymization

Project Examples

- Write a survey structured around a policy or regulation question (what should policy makers understand about AI in writing regulations?)
  - Some starting points:
    - AI Bill of Rights: https://www.whitehouse.gov/ostp/ai-bill-of-rights/
    - https://www.vox.com/future-perfect/2023/7/3/23779794/artificial-intelligence-regulation-ai-risk-congress-sam-altman-chatgpt-openai
    - https://techpolicy.press/

Project Examples

- Conduct a manual or automated meta-analysis of research processes in papers or other avenues where research discussions take place (social media, news)
  - Birhane, Abeba, et al. "The values encoded in machine learning research." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022
  - Leah Hope Ajmani, Stevie Chancellor, Bijal Mehta, Casey Fiesler, Michael Zimmer, and Munmun De Choudhury. *A Systematic Review of Ethics Disclosures in Predictive Mental Health Research*. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023
  - Media hype around AI
  - Prevalence of corporate research at different ML venues
  - Policy discussions around AI

Ideas for data sources

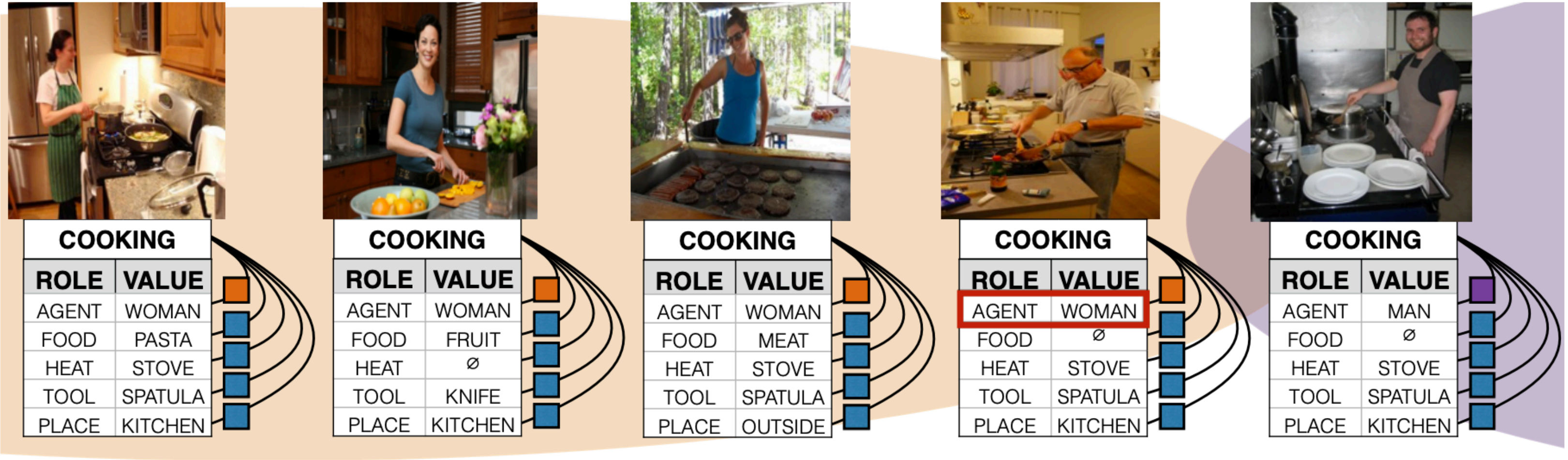
- Generally, data that has been used in prior research
  - Identify from reading research papers, which may have directly released the data or cited how they collected it (for example, this paper has a pointer to this archive of US Congressional Records). Authors are also often willing to share data if you reach out to them, even if they didn't post it publicly.
  - Shared tasks associated with workshops. For example, past versions of the workshop on NLP for Internet Freedom has had shared tasks on misinformation and censorship
  - Benchmark/shared task data is useful as analysis data, not just as model training data, for example this paper and this paper criticize standard NLP benchmarks
- There are some tools for data collection
  - Semantic Scholar API provides data on research publications (more details in the paper), which could be used for a variety of meta-analyses, like this paper on Big Tech influence in research
  - Many websites are possible to scrape (if you pay attention to terms of service and rate limits)
    - E.g. Wikipedia is a great source of data, and there are some existing archives, e.g. https://anjalief.github.io/wikipedia_bias_viz/
- There are some pre-collected archives of data
  - Common Crawl https://commoncrawl.org/ - large archive of web data
  - Twitter releases archives of data they've identified as potential information manipulation operations. It looks like they have not taken this down: https://transparency.twitter.com/en/reports/moderation-research.html

Today’s readings

- Classification (as opposed to generation – Thursday)

- Why did I assign these two particular papers on classification?
  - Extremely influential
    - Zhao et al. 2017: won EMNLP Best Paper
    - Obermeyer et al. 2019: referenced in pretty much every discussion of AI healthcare equity (and very often in AI equity settings)
  - Specific takeaways

Zhao et al. 2017: Bias *Amplification*



COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

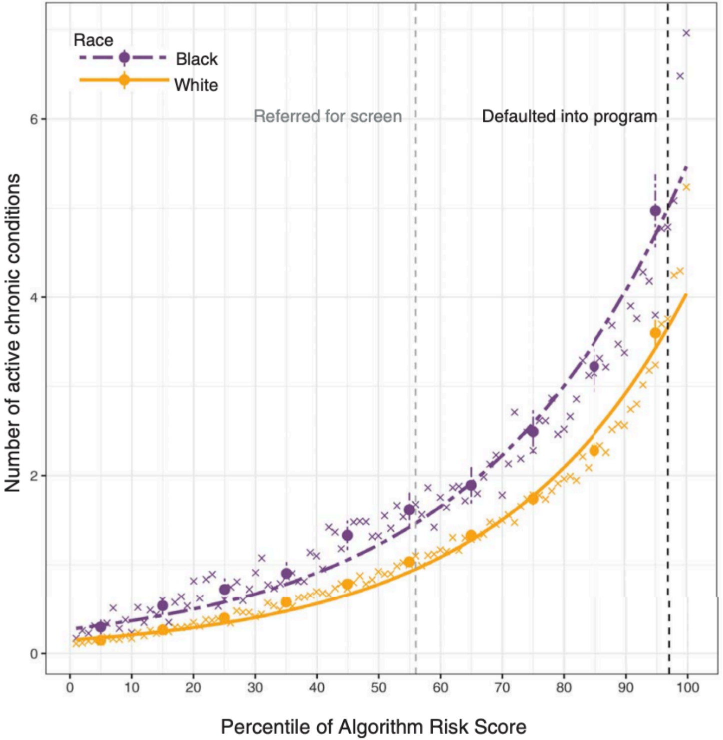
- "In the imSitu training set, **33%** of cooking images have man in the agent role while the rest have woman. After training a Conditional Random Field (CRF), bias is amplified: man fills **16%** of agent roles in cooking images"

Zhao et al. 2017: Bias *Amplification*

- Common confusion: if we correct for bias, do we risk “over-correcting”? E.g. might women’s products be advertised to men?
  - [Explicit gender-based targeting is a little different]
  - In this paper the motivating example is that “cooking” is under-associated with men: if a man has an image of cooking in his profile picture, he may be incorrectly shown ads for women’s products

- Bias *amplification* is (usually) more straightforward than other metrics
  - Example: ~90% of CEOs are men
    - If we ask an AI to generate an image of a CEO, we might disagree over whether it should be a man 90% of the time or 50% of the time, but we can probably agree that 99% of the time is wrong

Obermeyer et al. 2019: Choice of label matters



- Black patients are more sick than white patients at the same level of predicted risk
  - Problem originates in choice of label: model predicts healthcare costs, but black patients are less likely to receive treatment for the same ailments
- ~~Models are biased because data is biased~~
  - Zhao et al. 2017: models amplify bias
  - Obermeyer et al. 2019: practitioner choice of implementation (which label to predict) introduces bias

Discussion

Thursday 9/18

1. Feng et al. "From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models", ACL 2023.
2. Bianchi, Federico, et al. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale." FAccT 2023.
3. (optional) Myra Cheng, Esin Durmus, and Dan Jurafsky. "Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models", ACL 2023.