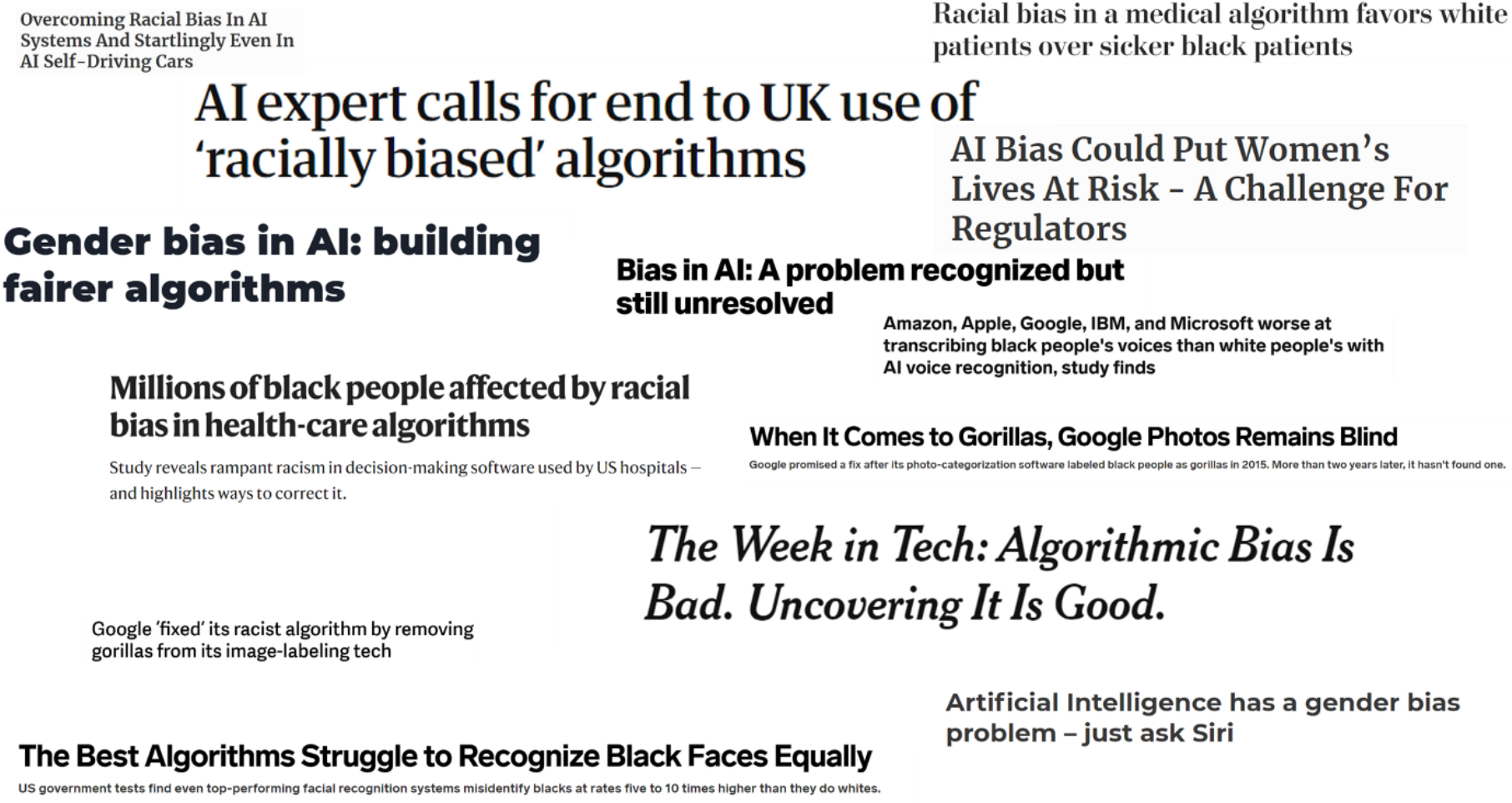
Privacy Risk

TECHNISCHE UNIVERSITÄT DARMSTADT



Source: https://towardsdatascience.com/algorithm-bias-in-artificial-intelligence-needs-to-be-discussed-and-addressed-8d369d675a70

Homework 4: Privacy

TECHNISCHE UNIVERSITÄT DARMSTADT

- Problem: Online data major source of training data for NLP
  - Can lead to detection of demographic characteristics
- The Dataset for this homework: Reddit post
  - Subreddits „funny“ and „relationships“

op_id	op_gender	post_id	post_text	subreddit
Osiris32	M	726999	Effort? Nevermind. Kidding, Im pretty sure you guys can help me with my Russian studies.	funny
Noble_toaster	M	1203690	Is A rich or a generally more stable person than you (Probably because she stayed with him the first time all those years ago)? If thats true then having a kid changes the whole situation. Shes got a kid now so she wants the practical dude. I dont doubt that she loved you more but shes not going back to you if she keeps this kid. Hell, you shouldnt even want her. Shell probably cheat on you sooner or later too. If the kid is yours and you want to be a father, fight for all your rights. Dont let her or A trample all over them. Get shared custody and all that.	relationships
argv_minus_one	M	722957	So do I. Not gonna hold it against her.	funny
Sete_Sois	M	1193116	Run to a galaxy far far away	relationships
courtFTW	W	741637	O Cristo Redentor is such a dope statue, I want to see it in person someday so badly!	funny
damiana8	W	1215679	Your bf is spineless and disrespectful, he is part of the problem if he wont stick up for you.	relationships

Homework 4: The provided Data

TECHNISCHE UNIVERSITÄT DARMSTADT

- subreddit_classifier.pickle — pretrained subreddit classifier
- gender_classifier.pickle — pretrained gender classifier
- test.csv — your primary test data
- male.txt — a list of words commonly used by men
  - Examples: Wife, girlfriend, beer, quality, burger, guys, business, casino
- female.txt — a list of words commonly used by women
  - Examples: husband, boyfriend, yummy, love, bf, sweet, hair, sister
- background.csv — additional Reddit posts that you may optionally use for training an obfuscation model

Homework 4: Tasks

TECHNISCHE UNIVERSITÄT DARMSTADT

1. **Baseline**:
   - On the test set the gender classifier achieves 64.6% and the subreddit classifier achieves 83.2% accuracy
   - **Your Task:** Obfuscate the data in the test.csv so that the classifier can not predict the gender of the authors while still being able to correctly predict the subreddit of a post.
2. **Obfuscation of the Test Dataset** (12P)
   1. Random (4P)

      For the posts written by men, replace appearing words from the male.txt randomly with one from the female.txt and vice versa
   2. Similarity (4P)

      Instead of random replace them based on semantic similarity (using cosine distance between pre-trained word embeddings)
   3. Your Idea (4P)

      Get creative and come up with your own idea (which can be related to the previous approaches from 1 and 2)

Homework 4: Tasks

TECHNISCHE UNIVERSITÄT DARMSTADT

3. **Advanced Obfuscated Model** (5P)
   - Develop your own classifier
   - Report baseline accuracy
   - Obfuscate train data and report changes
4. **Discussion: Ethical Implications** (3P)
   - What are demographic features (name at least three) and explain shortly some of the privacy violation risks?
   - Explain the cultural and social implications and their effects?
     In this context discuss the information privacy paradox. You may refer to a recent example like the COVID-19 pandemic
   - Name at least three privacy-preserving countermeasures.

Homework 4: Submission

TECHNISCHE UNIVERSITÄT DARMSTADT

- Submit your obfuscated versions of the test.csv (name them accordingly) for each task together with the .ipynb notebook as a zip-archive
- The notebook should contain your answers for Tasks 2-4:
  - Gender classification accuracy (2-3)
  - Subreddit classification accuracy (2-3)
  - Your brief commentary on the results in comparison to the other models and the baseline (2-3)
  - For task 3 shortly describe your model and train data modification choices
  - Answers to the discussion questions from task (4)
- Do not add other files
- IMPORTANT NOTE: Please use Google Colab to run the notebook.