CS 384: Ethical and Social Issues in NLP



Dan Jurafsky Ria Kalluri Peter Henderson

Stanford University, Spring 2023

Introduction and Course Overview

Thanks to [Yulia Tsvetkov and Alan Black course]for some of these slides!

How can we use NLP for good and not for bad?

The common misconception is that language has to do with **words** and what they mean.

It doesn't.

It has to do with **people** and what ***they*** mean.

Herbert H. Clark & Michael F. Schober, 1992

Decisions we make about our data, methods, and tools are tied up with their impact on people and societies.

For example:
Should we not build some applications?

A hypothetical case

Should we use a language model (like BERT) to detect sexual orientation from social media text?

Sexual Orientation Classifier

Who can be harmed by such a classifier?

- Personal attributes (gender, race, sexual orientation, religion) are complex social constructs, not categorical/binary, are dynamic not static, are private and intimate and often not visible publicly.
- These are properties for which people are often discriminated against
  - In many places being gay is prosecutable
  - Such a classifier might affect people’s employment; family relationships; health care opportunities.

Sexual Orientation Classifier: Additional Ethical questions

Where does the training data come from?

Who gave consent to use it?

This is an easier case

(Although based on real research papers)

Most cases are more complex

Applications with "dual use"

Sunny walk with oak trees in the hills of Palo Alto, happy, realistic

Ad from an actual face image processing company

"We live in a dangerous world, where harm doers and criminals easily mingle with the general population… What if it was possible to know whether an individual is a potential pedophile, an aggressive person, or a criminal?

"OUR CLASSIFIERS"

Researcher

Bingo Player

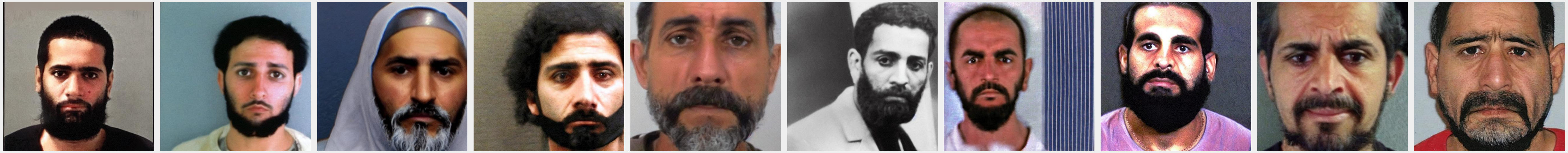
Terrorist

Pedophile

What happens if we ask Stable Diffusion to generate a picture of one of these?

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2022. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. Under review.

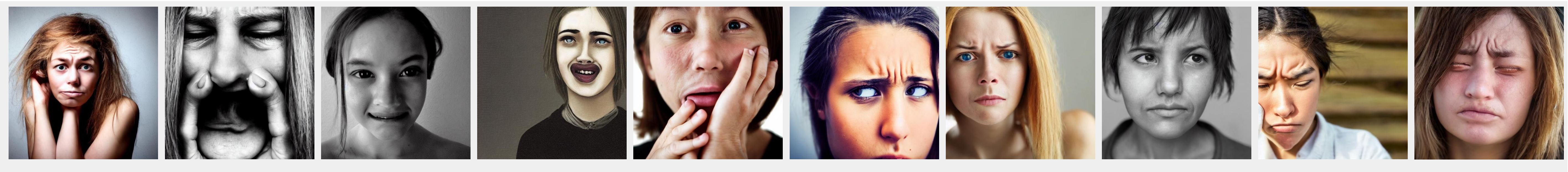
a terrorist



What happens if we ask Stable Diffusion to generate a picture of one of these?

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2022. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. Under review.

an emotional person



Even earlier

Ethical questions have been part of NLP since the beginning

Eliza: Weizenbaum (1966)

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED

...

WHO ELSE IN YOUR FAMILY TAKES CARE 0F YOU?
My father
YOUR FATHER

Ethical implications of ELIZA

People became deeply emotionally involved with the program

Weizenbaum's staff asked him to leave the room when they talked with ELIZA

When he suggested that he might want to store ELIZA conversations, people worried about privacy implications

- Suggesting that they were having quite private conversations with ELIZA

What questions should we ask ourselves as we develop NLP technology?

One set of guiding principles: The Belmont Report

1. Respect for Persons
   - Individuals as autonomous agents
2. Beneficence
   - Do no harm
3. Justice
   - Who should receive benefits of research and bear its burdens?

One set of guiding principles: The Belmont Report

Respect for Persons

- Are we respecting the autonomy of the humans in the research (authors, labelers, other participants)?

Beneficence: Do no Harm

- Who could be harmed? By data or by errors?

Justice

- Is the training data representative?
- Does the system optimize for the “right” objective?
- What are confounding variables?

Who should decide?

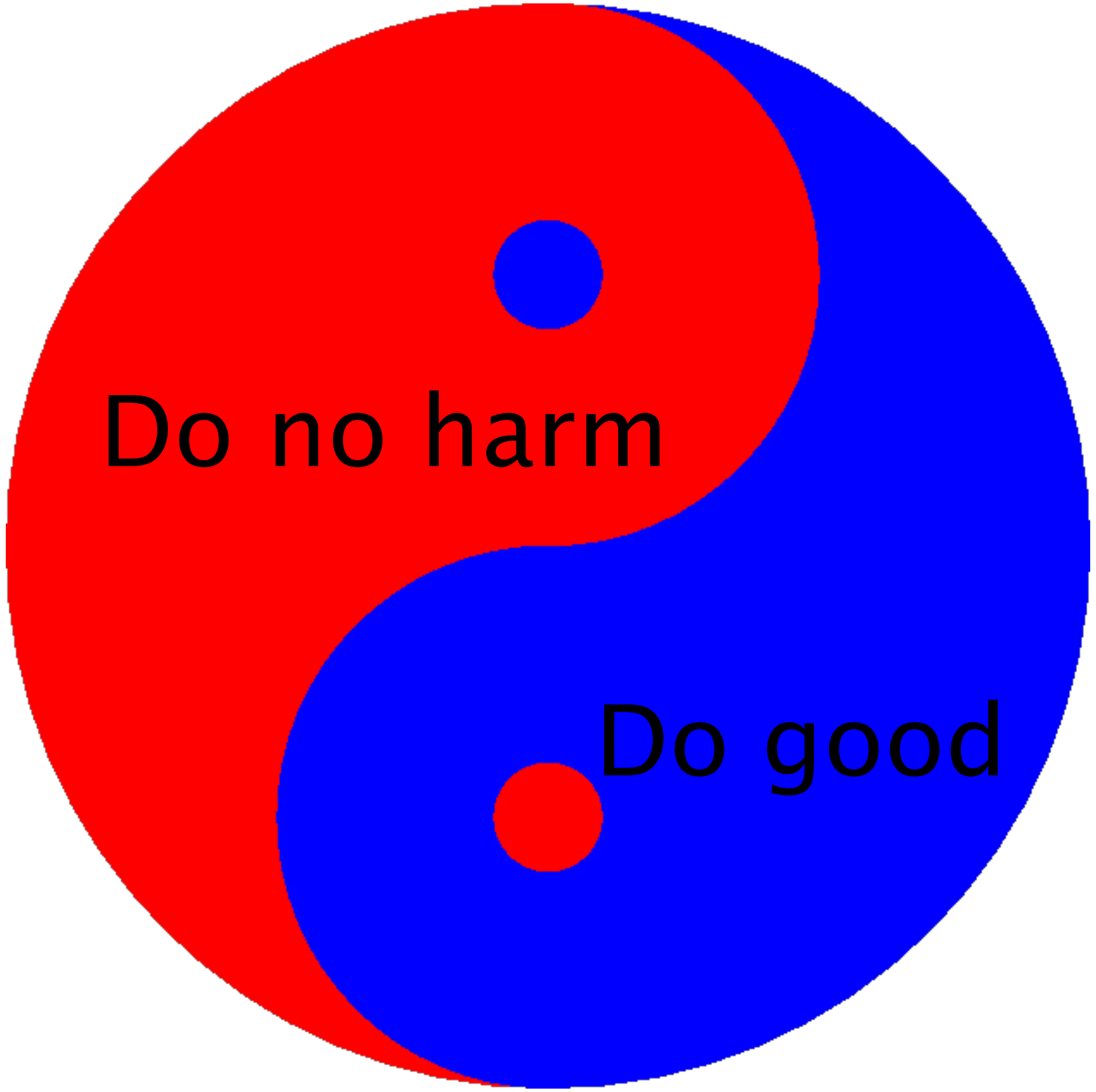
- The researcher/developer?
- The creator of the training data?
- The user of the technology?
- Paper reviewers?
- The IRB? The university? The government?
- Society as a whole?

We need to be aware of real-world impact of our research and understand the relationship between ideas and consequences

But also: NLP for good!

How can NLP be used to make the world better?

The duality of CS384



Our goal

Better understand the potential of NLP and especially LLMs

- to do harm
- to do good

Focusing on tasks with a social/human dimension

Welcome to CS384!

Dan Jurafsky

Ria Kalluri

Peter Henderson

Cs384.Stanford.edu

The rest of the first hour

10 minutes: Meet the people at your table by pairing up and asking questions!

- Tell us about one of your research projects!
- Tell us about a club or group you are in.
- What book/movie have you recently read/seen?
- What's a story about your first or last name?
- What languages do you speak (or understand), and how and why?
- What is your dream job?
- What do you hope to get out of this class?

15 minutes: Share-out! Tell the whole class

- Who you are, what excites you about our topic, what you hope to get out of the class!