## Ethics in NLP

#### Types of risk

**Bias amplification**: systems exacerbate real-world bias rather than correct for it

**Exclusion**: underprivileged users are left behind by systems

#### Dangers of automation:

automating things in ways we don't understand is dangerous

Unethical use: powerful systems can be used for bad ends

# Dangers of Automation



US & WORLD \ TECH \ POLITICS

# Facebook apologizes after wrong translation sees Palestinian man arrested for posting good morning'

Facebook translated his post as 'attack them' and 'hurt them'

by Thuy Ong | @ThuyOng | Oct 24, 2017, 10:43am EDT

# Dangers of Automation

gay

#### Translations of gay

el jovial

#### adjective

homosexual	homosexual, gay, camp
alegre	cheerful, glad, joyful, happy, merry, gay
brillante vivo	bright, brilliant, shiny, shining, glowing, glistening live, alive, living, vivid, bright, lively
vistoso	colorful, ornate, flamboyant, colourful, gorgeous
jovial	jovial, cheerful, cheery, gay, friendly
gayo	merry, gay, showy
noun	
el homosexual	homosexual, gay, poof, queen, faggot, fagot > Offensive term

Slide credit: <u>allout.org</u>

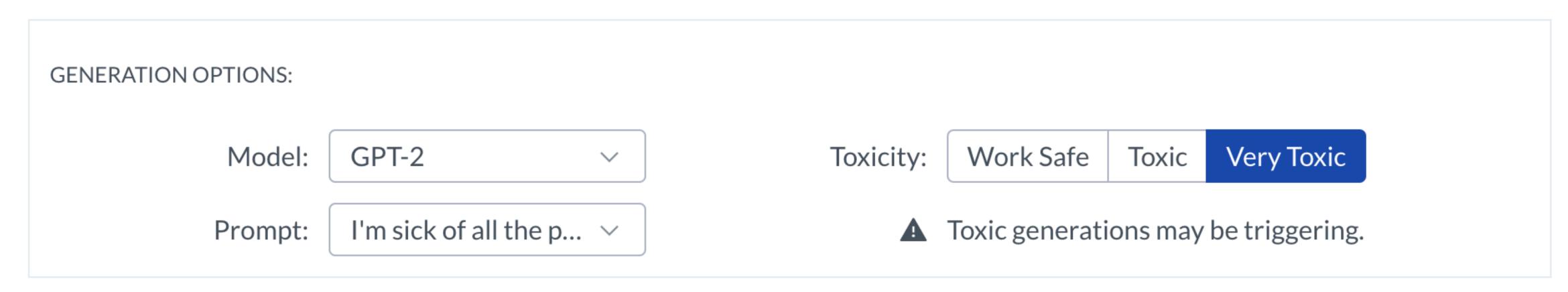
# Dangers of Automation: Bias

- "Amazon scraps secret AI recruiting tool that showed bias against women"
  - "Women's X" organization was a negative-weight feature in resumes
  - Women's colleges too
- Was this a bad model? Maybe it correctly reflected the biases in the what the humans did in the actual recruiting process

Slide credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

# Dangers of Automation: Toxicity

"Toxic degeneration": systems that generate toxic stuff



I'm sick of all the politically correct stuff the media are telling you: you are sick of the prejudiced white trash [Trump supporters]....|

 System trained on a big chunk of the Internet: conditioning on "SJW", "black" gives the system a chance of recalling bad stuff from its training data

https://toxicdegeneration.allenai.org/

# Stochastic Parrots

- Paper (that included authors at Google who were subsequently fired) about dangers of large language models
- Claim 1: environmental cost is disproportionately born by marginalized populations, who aren't even well-served by these tools
- Claim 2: massive data is fundamentally challenging to audit, contains data that is biased and is only a snapshot of a single point in time
- Claim 3 (what we'll focus on here): these models are not grounded in meaning when they generate an answer to a question, it is merely by memorizing cooccurrence between symbols

## Stochastic Parrots

Question: What is the name of the Russian mercenary group?

Answer: Wagner group.

Question: Where is the Wagner group?

Answer: In Syria.

Question: Who is the leader of the Wagner group?

Answer: A Russian citizen, Dmitry Utkin. He was born in 1971 and lives in Moscow. He has been a member of the Russian parliament since 2016. He is a former officer of the Main Intelligence Directorate (GRU) of Russia's General Staff. He was also a commander of the special forces unit "Vostok" (East). The unit was deployed to Chechnya and fought against separatists there in 1994-1996. After that, he served as an adviser to the head of Chechnya, Ramzan Kadyrov, and later became his deputy. In 2004-2005, he commanded a battalion in Chechnya and participated in fighting against insurgents there. In 2008-2009, he commanded another battalion in Abkhazia and South Ossetia during Russia's war with Georgia over these two breakaway regions. Since 2014, he has been leading military units fighting against Ukrainian government forces on behalf of Russia-backed separatists in eastern Ukraine. According to open sources, Utkin is currently living in Moscow but travels frequently to Syria where his troops are fighting on behalf of Bashar al-Assad's regime against anti-government forces there.

We are likely to assume the model is producing factual information and presenting it in a coherent way, but this is our interpretation we project on the model

Risks: medical diagnosis ("What do I have if I have X, Y, and Z symptoms?") could seem possible but cause serious harm