Social Computing Capstone

Day 15: Commercial Content Moderation

CSE 481p | Winter 2022

Amy X. Zhang
Assistant Professor | University of Washington, Allen School of Computer Science & Engineering

Schedule for today's class

- Share from A4 (Write a dystopia) (10 min)
- Discussion of reading and lecture on commercial content moderation (15 min)
- Group working time

Black Mirror writers' room

[Casey Fiesler 2021, tinyurl.com/blackmirrorwritersroom]

Share with the class:

Quickly tell us what your story was about.

- Is this something you could imagine happening today (or soon)?

- What's the underlying social anxiety or ethical issue your story is drawing out?

- How was the experience of writing a dystopia and reflecting on it?

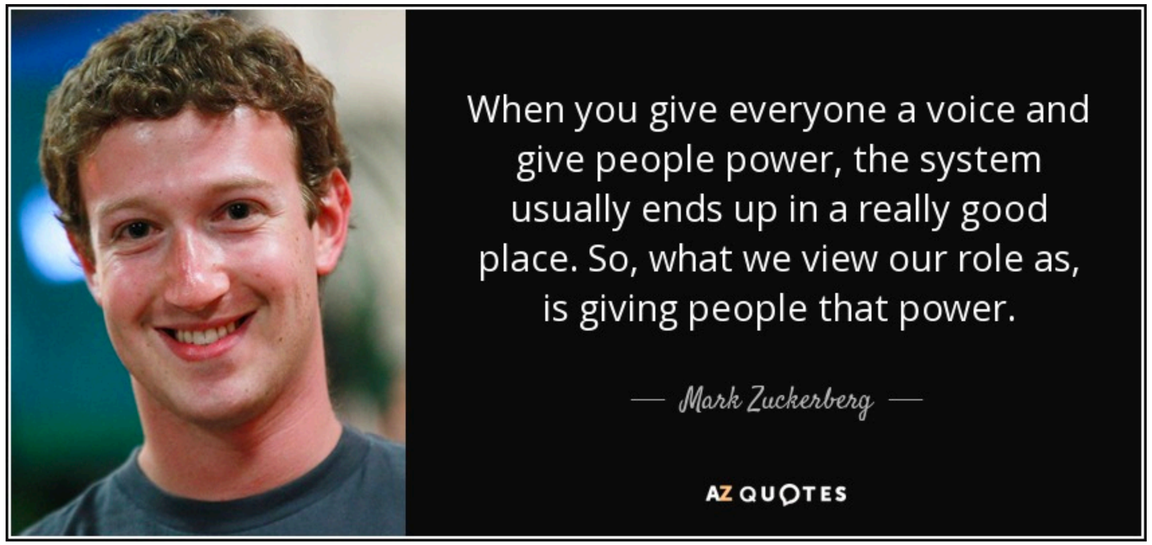
Commercial Content Moderation
(or alternatively, content moderation at scale)

Content moderation exists *everywhere* we have content

We have always had content moderation. Back when we consumed content via newspapers, TV, and radio, content moderation meant having editors and regulators overseeing journalists and producers. There were fewer avenues for media consumption, and “gatekeepers” were fewer and had more power.

Today, the nature of content moderation has changed drastically alongside the explosion of social media.

~ Flashback to the early-mid 2000's ~



"The fact that we're all connected…does change the parameters…But I think, in general, it's clear that most bad things come from misunderstanding, and communication is generally the way to resolve misunderstandings—and the Web's a form of communications—so it generally should be good…"

–Tim Berners Lee, inventor of the World Wide Web, 2006, interview

Social media today

How would you address some of the many problems we've discussed in class (harassment, misinformation, incitement to violence, threats, gore, bullying, etc….)?

Mark's answer: a large team of humans for now, and the hope of AI to eventually replace them

Why humans and why not AI today? Content moderation is really tricky!

- All nudity is banned.
- Actually, nude paintings and sculptures are ok.
- Actually, nude pictures of historical significance are ok.
- Actually, pictures showing genitalia in the context of birth or health-related photos are ok.
- Actually, female breasts are ok if they're breast-feeding.
- …Also if they are protesting.
- …Also if they are showing a post-mastectomy scar.

Fury over Facebook 'Napalm girl' censorship

By Zoe Kleinman
Technology reporter, BBC News

9 September 2016

Vietnam War

Support The Guardian
Available for everyone, funded by readers
Search jobs
Sign in
Contribute →
Subscribe →

News | Opinion | Sport | Culture | Lifestyle

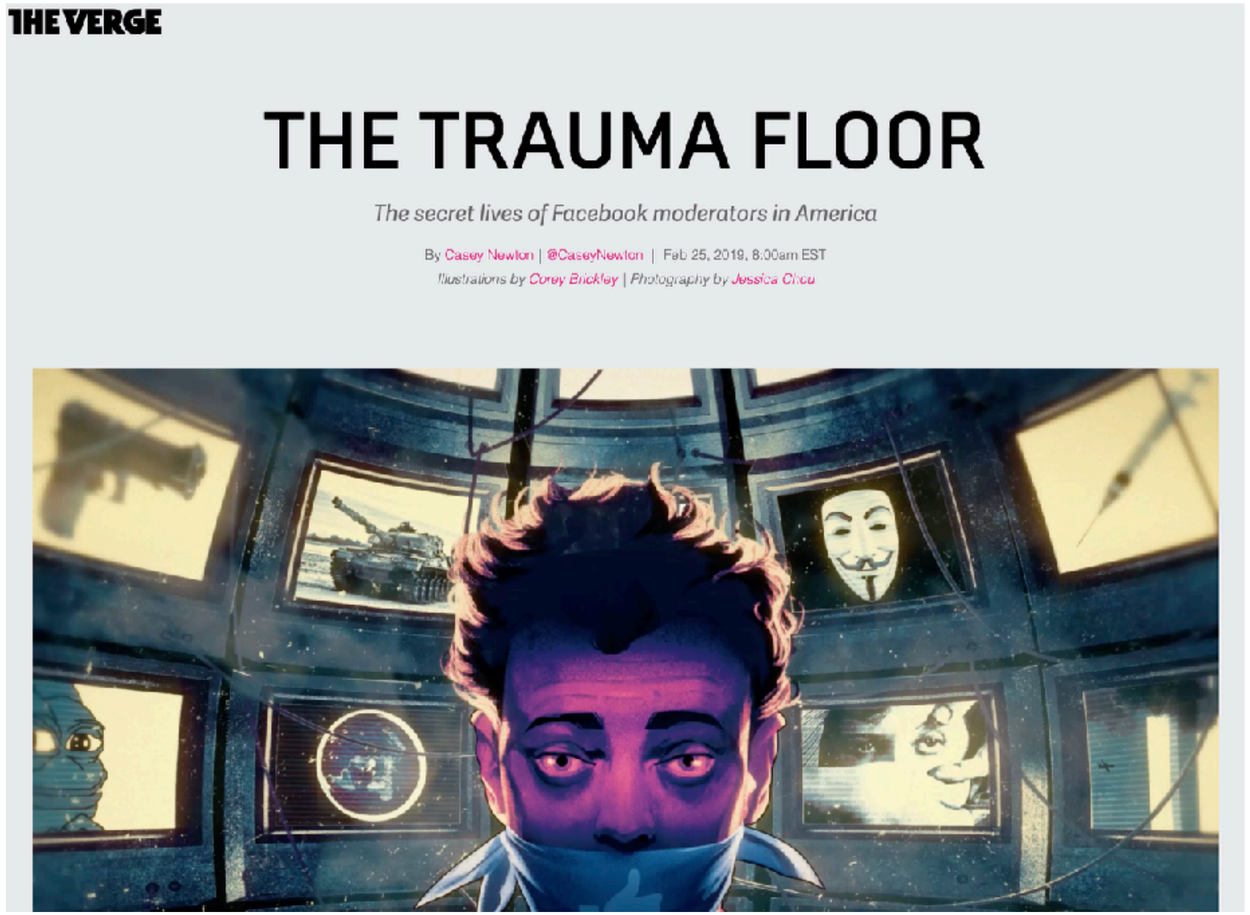
US Elections 2020 World Environment Soccer US Politics Business Tech Science Nev

Facebook

This article is more than 11 years old

Mums furious as Facebook removes breastfeeding photos

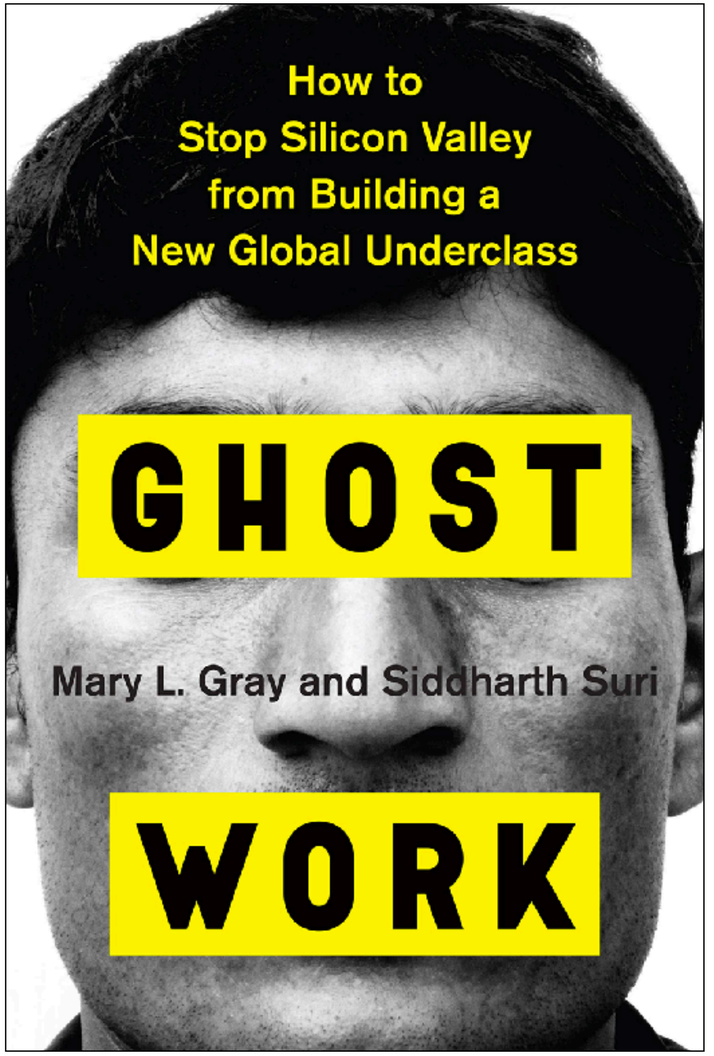
Commercial content moderation

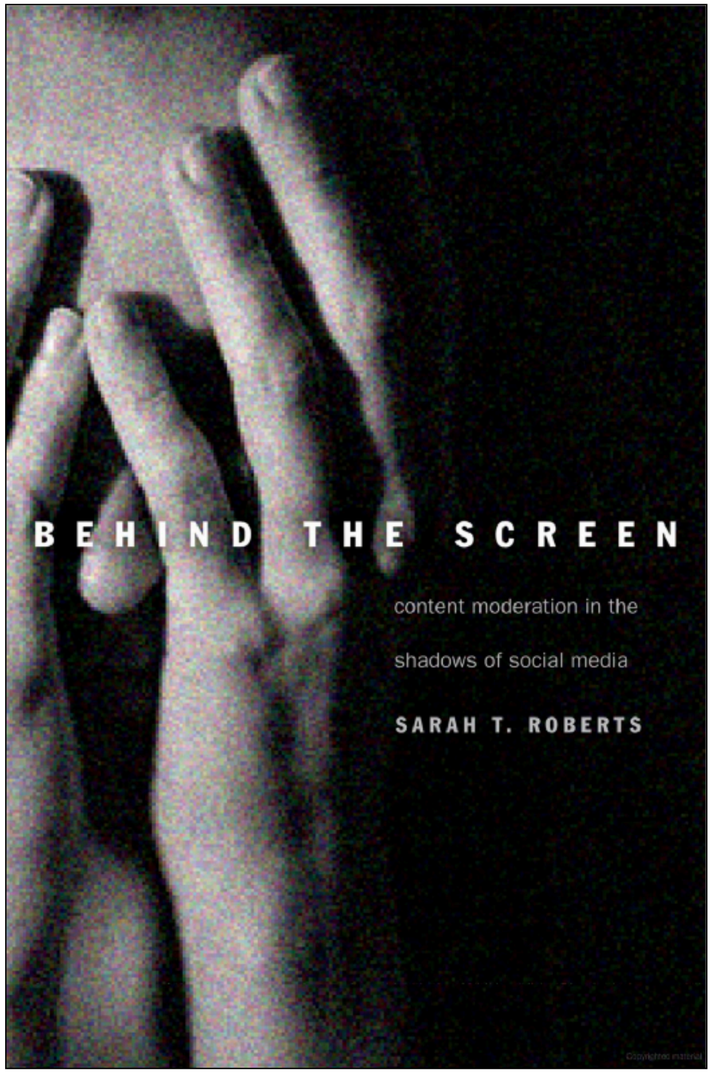


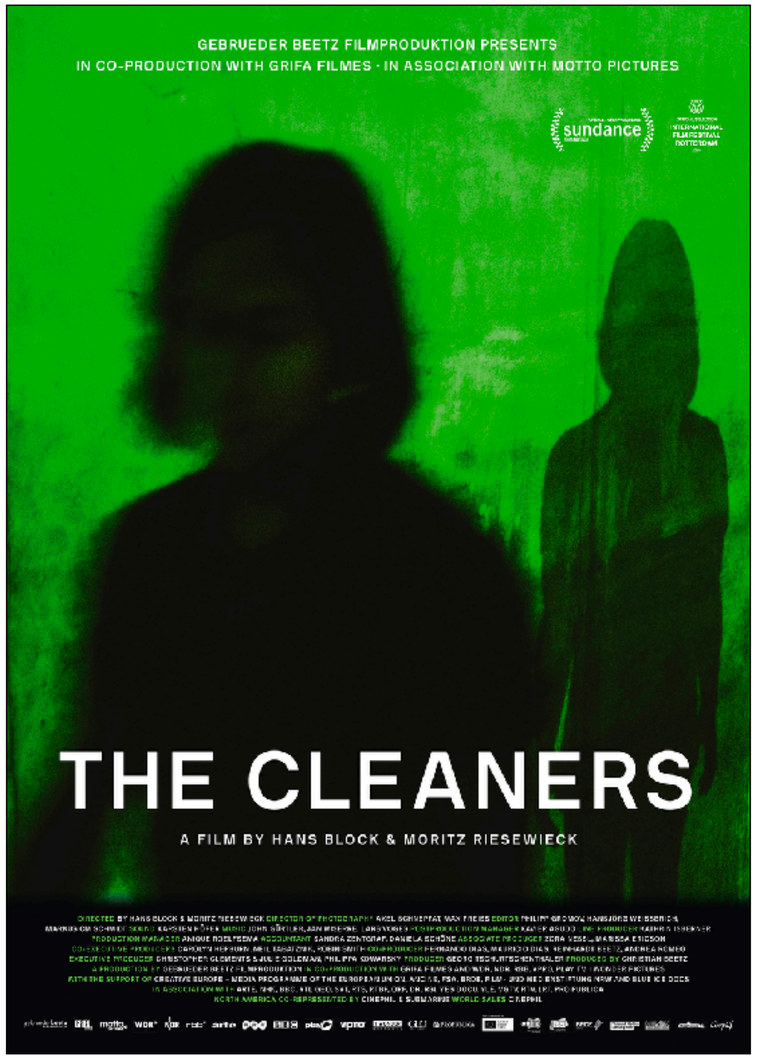
Just how many people? Facebook alone employs around 15,000 people. In total, the estimate is that around 150,000 “commercial content moderators” exist around the world.

Many are contractors, don’t have health benefits, see hundreds of pieces of content a day, taking <1 minute per piece, make close to minimum wage, and are outsourced to 3rd party companies in different places, in many cases completely different countries where wages are lower.

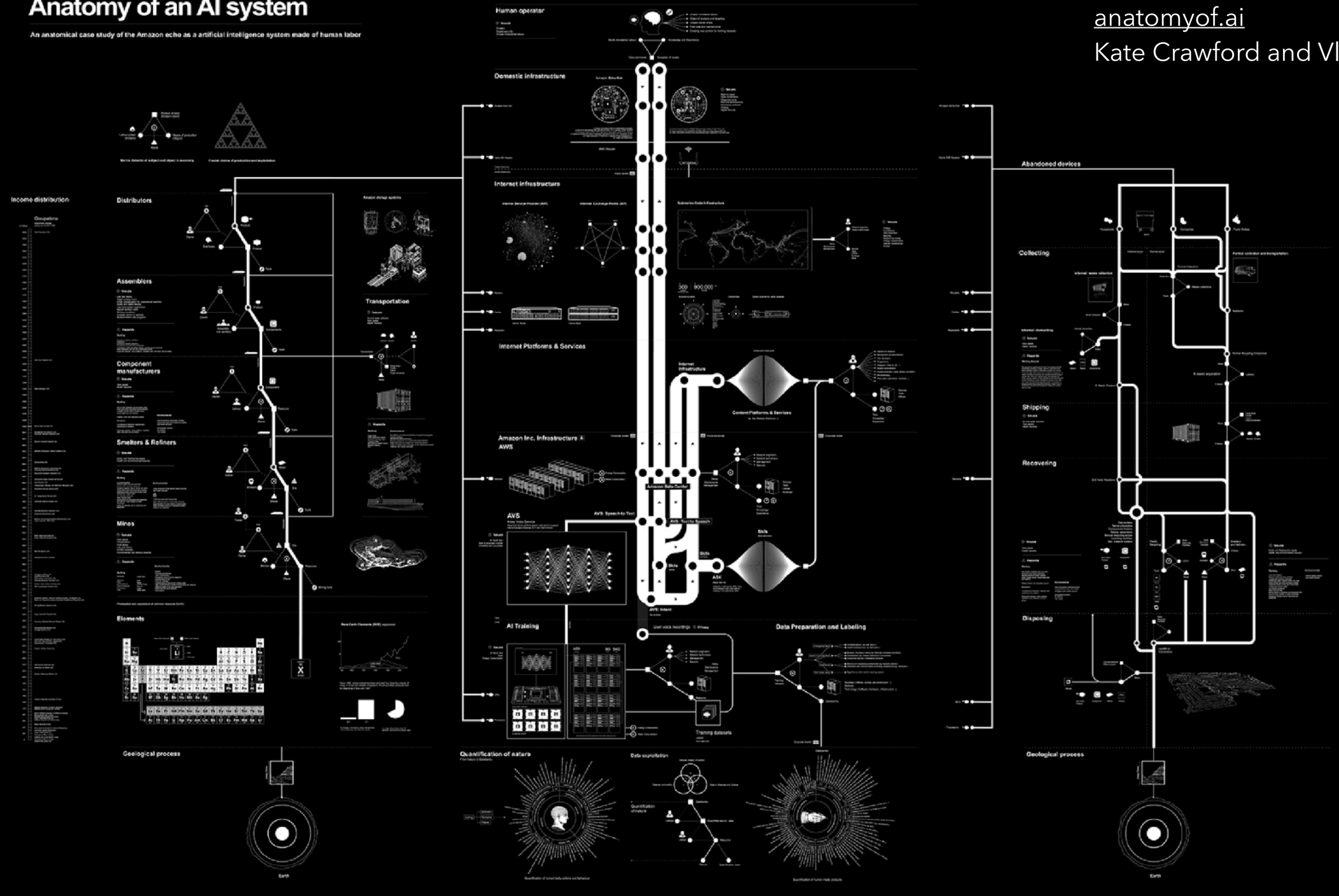
What are some potential problems with this setup?


How to
Stop Silicon Valley
from Building a
New Global Underclass
GHOST
WORK
Mary L. Gray and Siddharth Suri


BEHIND THE SCREEN
content moderation in the
shadows of social media
SARAH T. ROBERTS

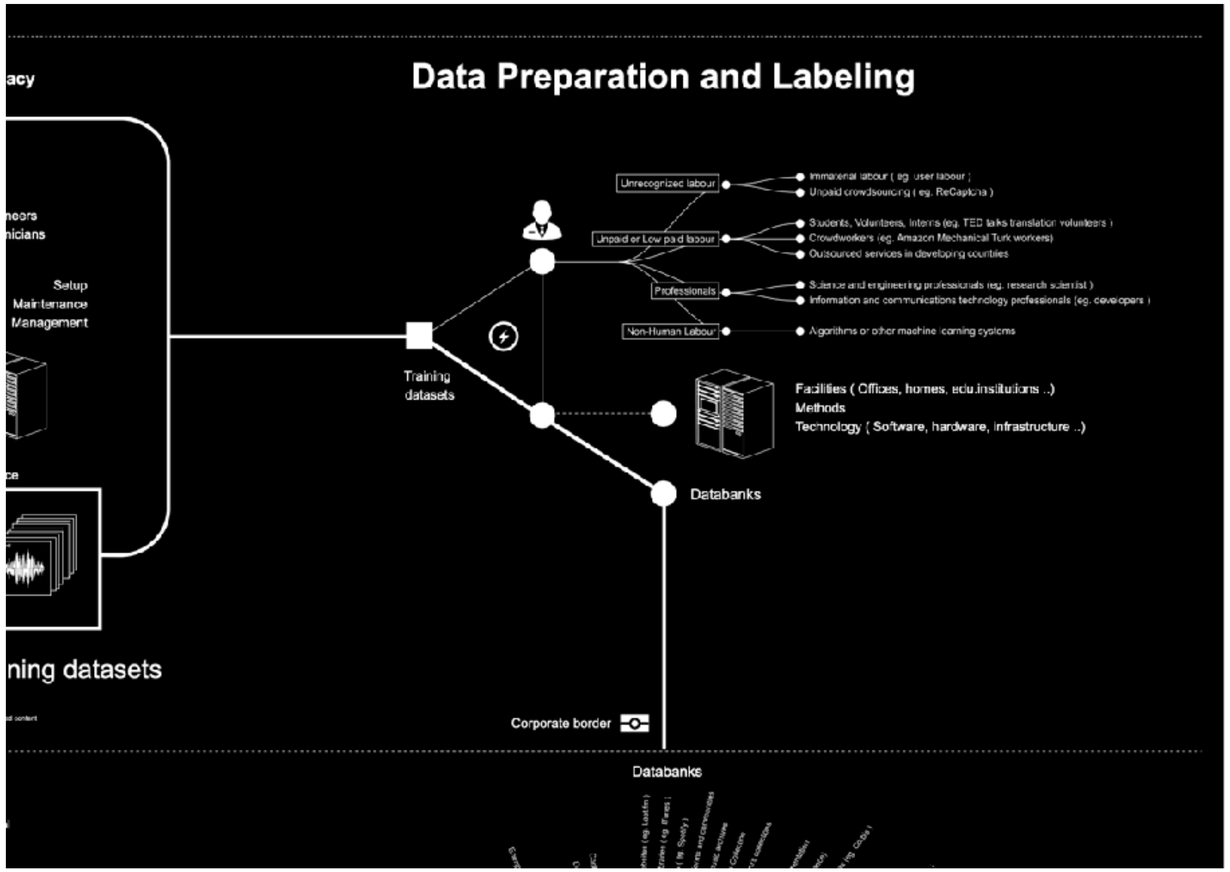

GEBRUEDER BEETZ FILMPRODUKTION PRESENTS
IN CO-PRODUCTION WITH GRIFA FILMES · IN ASSOCIATION WITH MOTTO PICTURES
sundance
THE CLEANERS
A FILM BY HANS BLOCK & MORITZ RIESEWIECK

AI content moderation doesn't mean there aren't humans



anatomyof.ai
Kate Crawford and Vladan Joler

Human resources behind AI



To train a model, you need lots and lots of data! To collect data, you need annotators to look at all that data and label it. So now you need to employ lots and lots of annotators. Sounds familiar?

There are also lots of cases where the model will be wrong. So you'll need a process where people are continually checking the output of the model and giving the correct label. Sounds familiar?

AI is biased…because people training them are biased

The Risk of Racial Bias in Hate Speech Detection

Maarten Sap[◇] **Dallas Card**[♣] **Saadia Gabriel**[◇] **Yejin Choi**[◇♡] **Noah A. Smith**[◇♡]
[◇]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA
[♣]Machine Learning Department, Carnegie Mellon University, Pittsburgh, USA
[♡]Allen Institute for Artificial Intelligence, Seattle, USA
msap@cs.washington.edu

Abstract

We investigate how annotators' insensitivity to differences in dialect can lead to racial bias in automatic hate speech detection models, potentially amplifying harm against minority populations. We first uncover unexpected correlations between surface markers of African American English (AAE) and ratings of toxicity in several widely-used hate speech datasets. Then, we show that models trained on these corpora acquire and propagate these biases, such that AAE tweets and tweets by self-identified African Americans are up to two times more likely to be labelled as offensive compared to others. Finally, we propose dialect and race priming as ways to reduce the racial bias in annotation, showing that when annotators are made explicitly aware of an AAE tweet's dialect they are significantly less likely to label the tweet as offensive.

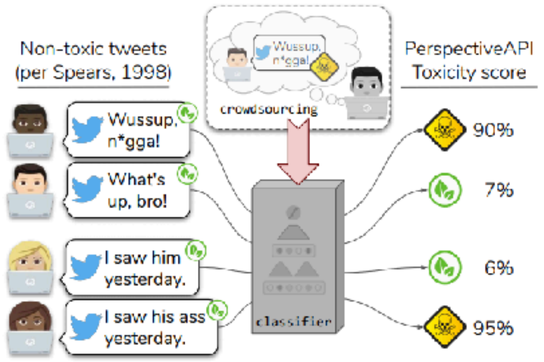


Figure 1: Phrases in African American English (AAE), their non-AAE equivalents (from Spears, 1998), and toxicity scores from PerspectiveAPI.com. Perspective is a tool from Jigsaw/Alphabet that uses a convolutional neural network to detect toxic language, trained on crowdsourced data where annotators were asked to label the toxicity of text without metadata.

You're never done. Data needs to be continually labeled

There are trillions of searches on Google every year. In fact, 15 percent of searches we see every day are new—which means there's always more work for us to do to present people with the best answers to their queries from a wide variety of legitimate sources. While our search results will never be perfect, we're as committed as always to preserving your trust and to ensuring our products continue to be useful for everyone.

Takeaway: consider the well-being of your human moderators/annotators

You will always need humans involved in content moderation. The perfect AI doesn't exist for these kinds of tasks, and even if you do have a decent one (trained on labels by humans), it will constantly need to be updated with more data.

How can policies and technologies better support the well-being of these people?

- mental health support, better wages, less-precarious employment, ability for advancement, empowerment in their role as experts, visibility on the platform, supporting a local community instead of "view from nowhere"
- tools targeted to classify the worst content so people don't have to see it? or blur out parts that still capture the gist? or find near duplicates so things need only be labeled once? Connection to local child/animal support services to find the sources of this content?