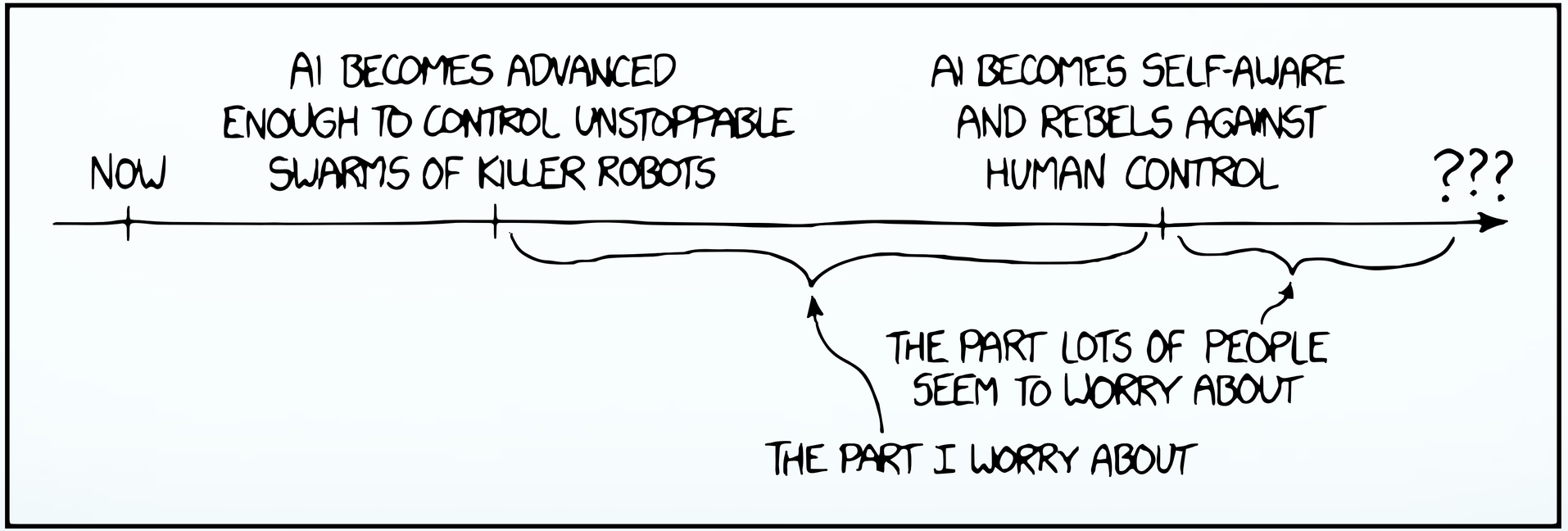
Ethics in NLP: Including Society in Discourse and Design

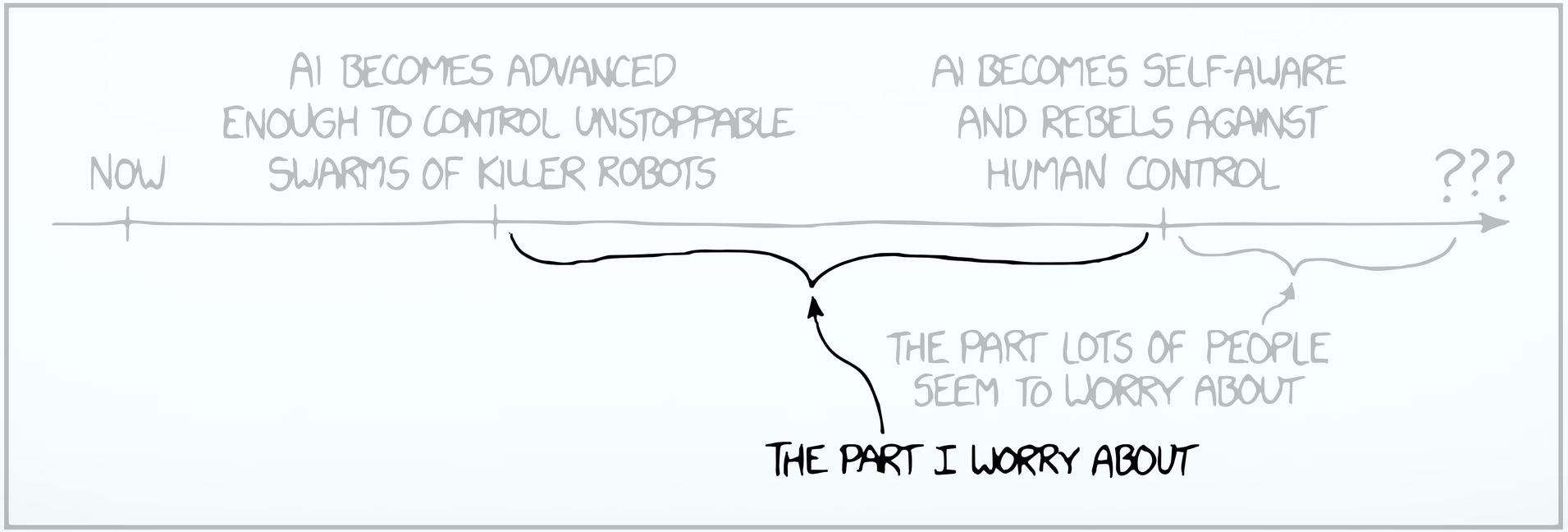
January 10th, 2019

Ryan Georgi

In A Nutshell… xkcd.com/1968



In A Nutshell… xkcd.com/1968



Alt Text: *"I mean, we already live in a world of flying robots killing people. I don't worry about how powerful the machines are, I worry about who the machines give power to."*

A Brief History of Research Ethics

Human Subjects Protections: A Brief History

- Much of current human subjects protections stem from ***Nuremberg Code***
  - Nuremberg Code arose from the Nuremberg Medical Trial, 1947
  - Physicians had experimented on prisoners

Nuremberg Code [wikipedia][BMJ]

1. ***Required is the voluntary, well-informed, understanding consent of the human subject in a full legal capacity***.
2. The experiment should aim at positive results for society that cannot be procured in some other way.
3. It should be based on previous knowledge (e.g., an expectation derived from animal experiments) that justifies the experiment.
4. The experiment should be set up in a way that avoids unnecessary physical and mental suffering and injuries.
5. It should not be conducted when there is any reason to believe that it implies a risk of death or disabling injury.

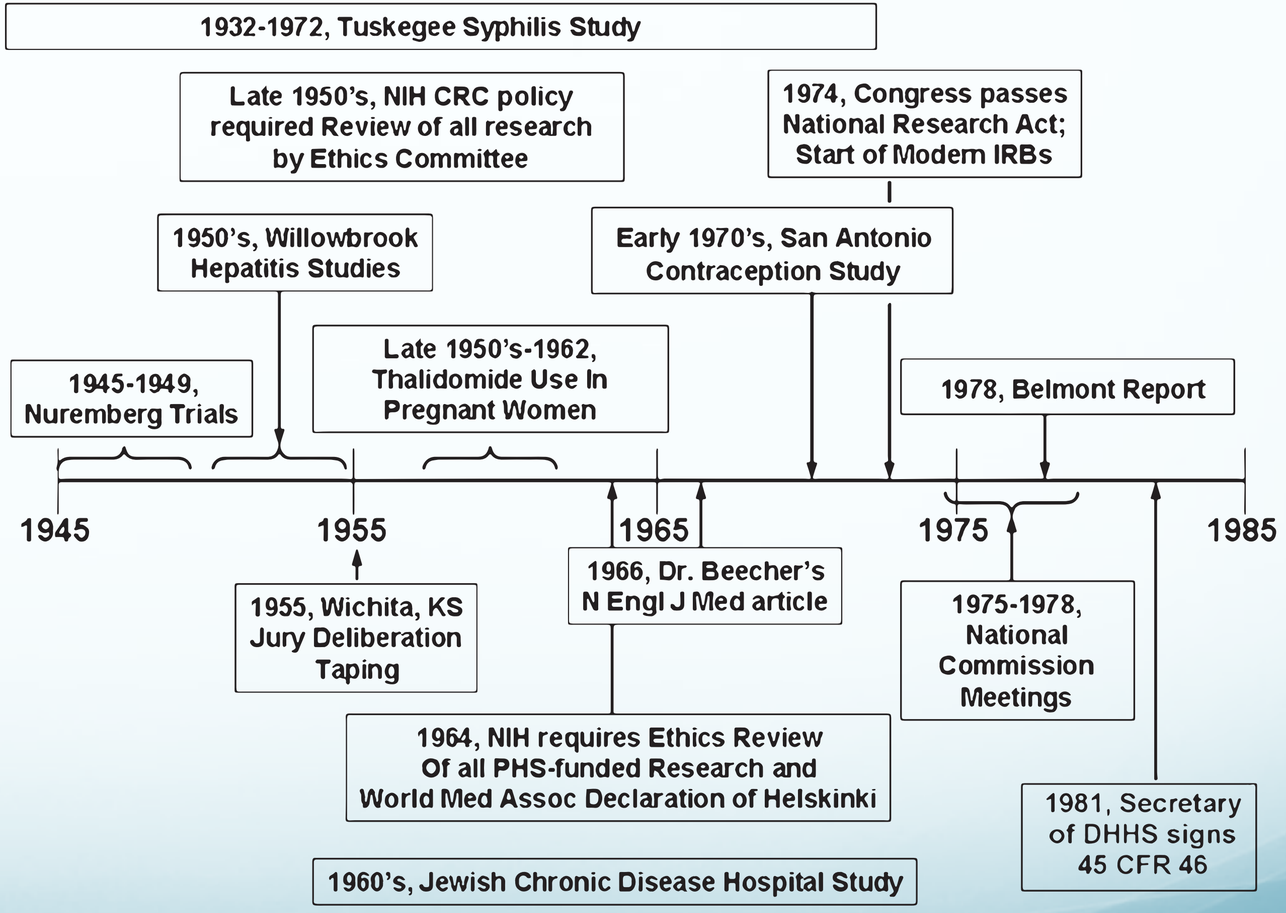
Nuremberg Code

6. ***The risks of the experiment should be in proportion to (that is, not exceed) the expected humanitarian benefits***.
7. Preparations and facilities must be provided that adequately protect the subjects against the experiment's risks.
8. The staff who conduct or take part in the experiment must be fully trained and scientifically qualified.
9. ***The human subjects must be free to immediately quit the experiment at any point when they feel physically or mentally unable to go on***.
10. Likewise, the medical staff must stop the experiment at any point when they observe that continuation would be dangerous.

Human Subject Abuses: A Timeline

(aka *Fig.1: Why We Can't Have Nice Things*)

- Timeline in the development of regulations on human-subjects research protections and institutional review boards (IRBs).
  - **NIH**: National Institutes of Health.
  - **CRC**: Clinical Research Center
  - **PHS**: Public Health Service.
  - **DHHS**: Department of Health and Human Services.
  - **CFR**: Code of Federal Regulations



[via Rice, 2008]

Human Subject Abuses: A Brief History

[via Rice, 2008]

- **Wichita Kansas Jury Taping (1955)** — Nonconsensual audio recording
- **Tuskegee Syphilis Study (1932–1972)**
  - Non-treatment of syphilis in Black Americans
  - Participants not informed they were not receiving care
- **Thalidomide in Pregnant Women (1950s-62)** — Participants not informed of exper. nature
- **San Antonio Contraception Study (1970s)** — Blind placebo study, no informed consent
- **Stanford Prison Experiment (1971)**
  - Psychological abuse, participants not allowed to withdraw consent
- …then I got really sad making this list, so let's move on.

Human Subjects Protections:
Reactions to Abuses

- **The National Research Act of 1974**
  - Largely spurred by furor over Tuskegee syphilis study
  - Established modern IRB system
  - Formed the National Commission
    - (for Protection of Human Subjects of Biomedical and Behavioral Research)
- **The Belmont Report (1978)** [HHS Website]
  - Product of National Commission
  - Three Principles — (1) **Respect for Persons** (2) **Beneficence** (3) **Justice**

The Belmont Report Principles:

Respect for Persons

1. Participants must ***voluntarily consent*** to participate in research.
2. The consent must be ***informed consent***.
3. Participants' ***privacy and confidentiality*** must be protected
4. Participants have the ***right to withdraw*** from research participation ***at any time*** without penalty.

The Belmont Report Principles:

Beneficence

- Research must ***maximize benefit*** and ***minimize harm***.
- Any risks must be justified by benefits to society or individual.
  - This is difficult, made on case-by-case basis

The Belmont Report Principles:
Justice

- Research should ***not*** systematically select specific classes of individuals who:
  - Are readily available because of where the research is conducted
  - Who are "*easy to manipulate as a result of their illness or socioeconomic condition*."
- Enrollment should reflect research objectives
  - Not be overly broad
  - Contemporary discussion
    - …***must include*** people who might be/are affected?

The Belmont Report Principles:
Justice

- Research should ***not*** systematically select specific classes of individuals who:
  - Are readily available because of where the research is conducted
  - Who are "*easy to manipulate as a result of their illness or socioeconomic condition.*"
- Enrollment should reflect research objectives
  - Not be overly broad
  - Contemporary discussion
    - …***must include*** people who might be/are affected?

Keep this in mind when discussing "existing data"

Something to Consider

- Though the Nuremberg code is most commonly cited
  - Weimar German government actually issued guidelines in 1931:
    - Guidelines for Human Experimentation
- Not the case that bad people did things that noone foresaw, ***then*** rules followed

Something to Consider

- Coming up with rules is the easy part
- Harder is designing robust systems to:
  - Require effective training
  - Prevent violations, intentional or accidental
  - Enforce compliance, and professional consequences for violations

Applications to NLP

Applications to NLP

- **Systems and Regulations for NLP Researchers**
  - The Lack of Data Science IRBs
  - Professional Code of Conduct?
- **Ethical Experimental Design**
  - Informed Consent?
  - Study of Beneficience vs. Harm?
  - Algorithmic fairness
- **How to communicate research findings?**

The Lack of IRBs for Data Science

via UW's Human Subjects Exemption Worksheet

☐ Category 4. Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.

The phrase "if these sources are publicly available" has been removed by the UW from the regulatory description of Category 4, because public data cannot be private, and if there is no collection of private identifiable data or interaction with subjects, there is no involvement of human subjects.

The Lack of IRBs for Data Science

- But…

- **Private information** includes information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place and information that has been provided for specific purposes by an individual and that the individual can reasonably expect will not be made public (e.g., a medical record).

- Now, this would seem to easily preclude Twitter, right?
  - But "*an individual can reasonably expect*"
  - So, we are using a "reasonable person" standard… so let's ask people.

Users' Expectations in Online Research

- Sociological Study — Williams et. al (2017)
  - Asked users a number of questions about their expectations for use of Twitter with respect to research

Expect to be asked for consent

Disagree	7.2	20.3
Tend to Disagree	13.1	
Tend to Agree	24.7	79.7
Agree	55	

Expect to be anonymized

Disagree	5.1	9.9
Tend to Disagree	4.8	
Tend to Agree	13.7	90.1
Agree	76.4	

Users' Expectations in Online Research

- Wide majority of users expect some degree of privacy in their Twitter usage
  - What role does Twitter's ToS play?
- This isn't exactly new
  - Having a discussion with a friend in a loud bar
    - Is it private? ***No.***
    - Do you expect to be recorded? ***Definitely not.***

Users' Expectations in Online Research

- Problem of "publicly available" stems from older definitions
  - Previously, being published meant you had to go out of your way
    - Talk to a reporter
    - Write a letter to the editor or op/ed
- Current usage does not sync with the majority of people's understandings now

Users' Expectations in Online Research

- Most online research at odds with principle of ***Respect for Persons***

What do we do next?

Professional Code of Conduct

Professional Code of Conduct

- A professional code of conduct could be one tool for ***enforcing compliance***
  - Lawyers can be disbarred
  - American Speech-Language-Hearing Association (ASHA) — sanction policies
    - Reprimand, Censure, Revocation of Membership, Withholding of Certification
  - IEEE also has CoE… since 1963 [current]
- ***…the Association for Computational Linguistics does not have such a code.***
  - Blog post from Hal Daumé [link]

Professional Code of Conduct

- Because no licenses issued by many tech societies, hard to sanction violators
- Comp Ling also much easier to practice in private industry
  - How to address this?

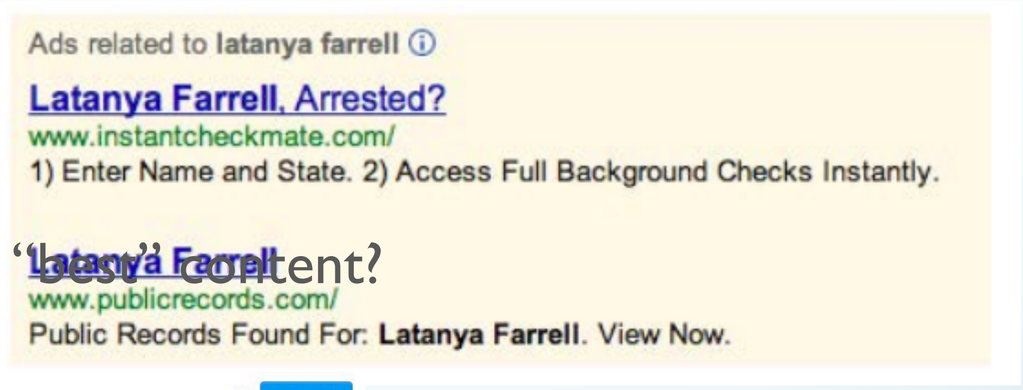
Harm & Bias

Dimensions of Harm & Bias

- **Acute Harm**
  - The impacts of unethical professional conduct are *acutely* visible
  - *Personal* level — can be easy to understand the impacts on a peer
  - (See: Slate Article on University of Rochester abuse scandal **Trigger warning for psychological abuse*)
- **Diffuse Harm**
  - Biased data or socially irresponsible system design
  - *Societal* level — can be difficult to understand the impacts on a hypothetical other
  - (See: Pro Publica's article on Broward County Recidivism algorithm, COMPAS)

Dimensions of Harm & Bias

- **Harms of Allocation**
  - How are home loan interest rates algorithmically determined?
  - How do admissions or hiring algorithms rank candidates?
- **Harms of Representation**
  - What do you get when you Google your name?
  - What do user engagement algorithms think is the "best" content?

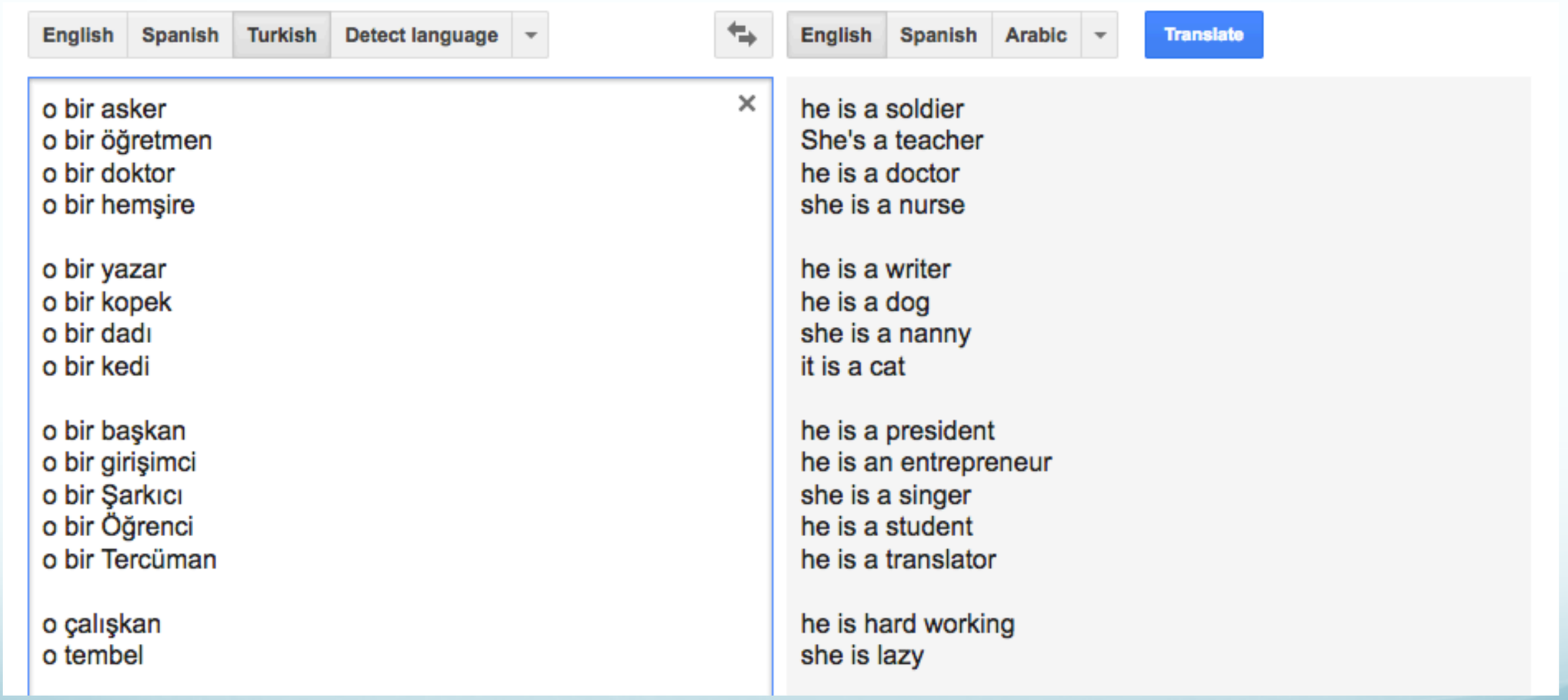


Positivity Break!

Bias:
Garbage In, Garbage Out

- Your training data has bias → your system will be biased.
  - Amazon built an AI to hire people, but reportedly had to shut it down because it was discriminating against women (Isobel Hamilton, *Business Insider*, 10/10/2018)
  - Gender and Dialect Bias in YouTube's Automatic Captions (Tatman, 2017)
  - Google Translate's gender bias pairs "he" with "hardworking" and "she" with lazy, and other examples (Nikhil Sonnad, *Quartz*, 11/29/2017)
  - Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (Bolukbasi et. al, 2016)

Language Data as Ground Truth



Algorithmic Fairness

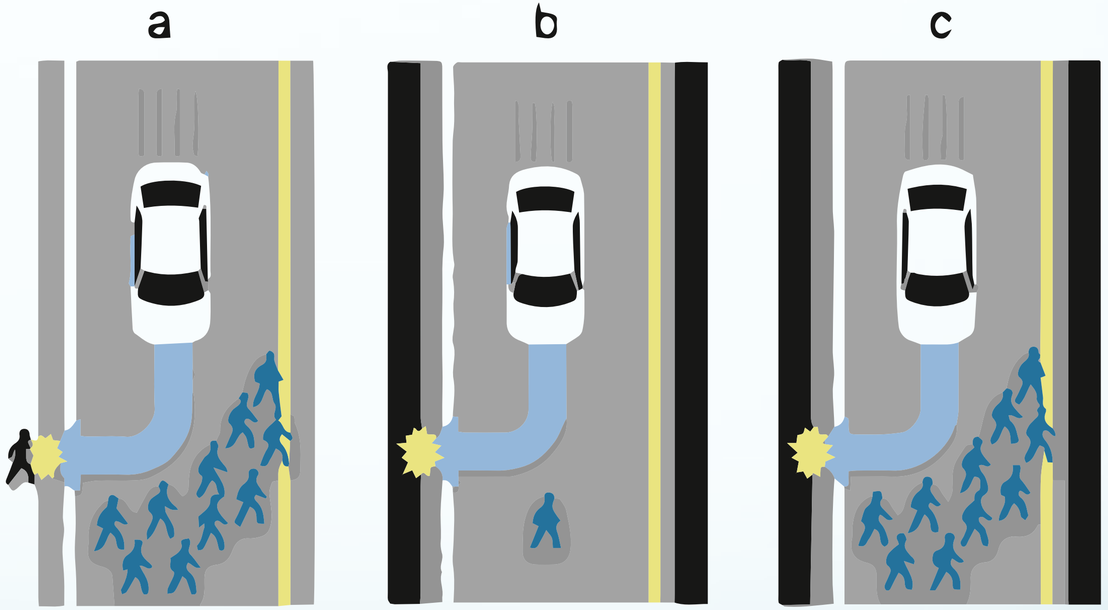
- Mathematically, however, bias is tricky!
  - To take one simple example, suppose we want to determine the risk that a person is a carrier for a disease X, and suppose that a higher fraction of women than men are carriers. Then our results imply that in any test designed to estimate the probability that someone is a carrier of X, at least one of the following undesirable properties must hold:
    a. the test's probability estimates are systematically skewed upward or downward for at least one gender; or
    b. the test assigns a higher average risk estimate to healthy people (non-carriers) in one gender than the other; or
    c. the test assigns a higher average risk estimate to carriers of the disease in one gender than the other.
    - The point is that this trade-off among (a), (b), and (c) is not a fact about medicine; it is simply a fact about risk estimates when the base rates differ between two groups. Kleinberg et. al (2016)

The Trolley Problem

- A thought experiment:
  - A runaway trolley is hurtling down a track, and there are five people at the end of the tracks.
  - If you are standing by a switch that will let you divert the trolley onto a second track where only one person stands, do you pull the lever?

The Trolley Problem

- The autonomous vehicle formulation:
  - Should the vehicle cause a collision that kills the driver to avoid killing multiple other people?
    - What about a pregnant woman? Child? etc



The Trolley Problem Problem

(The Trolley Metaproblem)

- What is the actual percentage of accidents in which this is a feasible solution?
  - According to IIHS-HLDI, 55% of fatal accidents are single-vehicle.
  - …of the 45% that were multiple vehicle, how many:
    - Weren't caused by driver error
    - Weren't caused by illegal operation
    - Weren't avoidable by braking sooner
- Ultimately, I'd be surprised if this was as high as 1% of just fatal accidents
  - At 1 death per 100 million miles (all fatal collisions)
  - This means 1 death per 100 billion miles (situations in which trolley problem applies)

The Trolley Problem Problem

(The Trolley Metaproblem)

- If our solution was correct 99.9999% of the time (0.0001% false positive rate)
  - We would potentially save 1 fatality per 100 billion miles…
  - …we would get 10,000 false positives per 100 billion miles (!!!!!!!!!!)
- ***and this is before taking into account programming bugs***
  - ***…which is a source for human error!***

Takeaways from the (Meta-)Trolley Problem

- **The false positive paradox**
  - When incidence of true event is less probable than the false positive rate
  - The test will give more false positives than true positives
  - …pay extremely close attention when modeling low probability events

- Utilitarian vs. Kantian/Deontological ethics

Kantian vs. Utilitarian Ethics

- **Kantian View**:
  - The rightness or wrongness of an action does not depend on the consequences.
  - "The Categorical Imperative"
- **Utilitarian View**:
  - We are responsible not just for our actions, but their consequences as well.
  - An individual negative action is preferable if consequence maximizes net "utility."
  - *"The greatest good for the greatest number."*

Applying Ethical Frameworks

- What are the ethical consequences of:
  - **True** Positive? **True** Negative?
  - **False** Positive? **False** Negative?
- In the US criminal justice system, for instance: (via O'Neill, 2016)
  - Presume innocence until proven guilty.
  - Burden of proof is to prove "beyond a reasonable doubt"
    - Suggested to be >95% confidence (Newman, 2007)
  - **Assumption**: it is far better to let a murderer walk free than send an innocent person to prison for life.

Applying Ethical Frameworks

- This is, in essence, handicapping the "efficiency" of the system in favor of fairness.
  - Deliberately letting the recall drop in favor of precision
  - Because we feel that more false positives are worse than a higher F_1-score
- How do we optimize for "fairness?"
  - Very open question
  - One suggestion (Hardt et al, 2016)
    - Make our models' ***error rates*** equivalent across different groups
    - Otherwise, varying incidence among different populations will result in FP/FN disparity

Some Example Systems in NLP

- Pre-emptively flag dating site users for likelihood to abuse (based on reported users model?)
  - + Protect users from abusive interactions
  - – What are the ethics of diagnosing someone without consent, or without a medical license?
- Identify symptoms of severe mental illness from clinical text (Jackson et al, 2016)
  - + Help healthcare professionals identify, intervene early
  - – Misdiagnosis, particularly if system relied upon too heavily
    - Drugs for dealing with mental health problems often powerful, have side effects
- Others…?

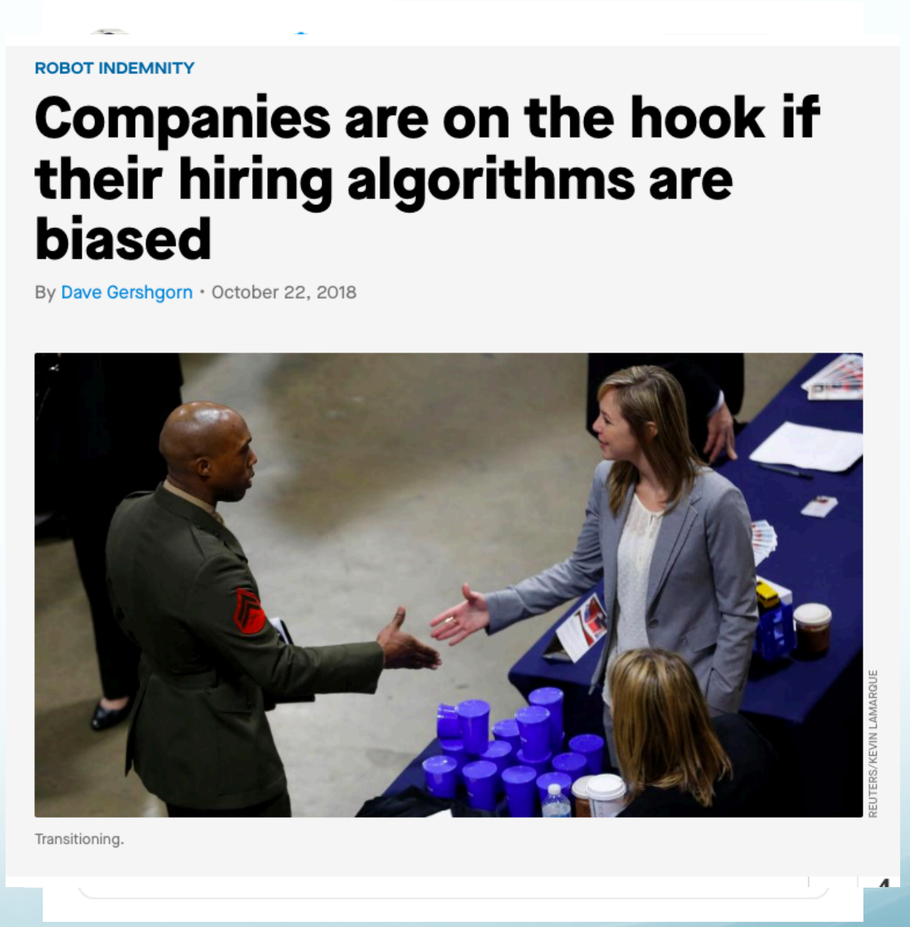
Some Example Systems in NLP

- A ranking system to prioritize callbacks for job candidates:
  - *Utilitarian*
    - It's okay to use protected class demographics system to promote more fair outcomes
  - *Deontological*
    - Using protected class information to make a decision is inherently problematic

Communicating our Research

Science Communication in NLP

- Problem — there are real risks in NLP research
  - …but these are not always what gets attention.
- How can we:
  1. Take the concerns of the public seriously?
  2. Be honest about the risks of technology?
  3. Mitigate the effects of hype and "puffery"?



ROBOT INDEMNITY

Companies are on the hook if their hiring algorithms are biased

By Dave Gershgorn • October 22, 2018

REUTERS/KEVIN LAMARQUE

Transitioning.

Science Communication

- "Because science communication seeks to inform decision making, it must begin by listening to its audience, to identify the decisions that its members face"

  (Fischhoff, 2013)

Lay Summaries & AI Safety Disclosures

- Proposal—Researchers should include both:
  - A "lay summary"
    - Accurate, but accessible description of the work for the general public
  - AI Safety Disclosure
    - List failure modes — both empirical and qualitative

Course Structure

Assignment Structure

- **15% — Weekly Readings and Discussion Participation**
  - This will largely be a reading-group style seminar
  - I will have a list of readings on the bibliography
  - Readings will typically be a choice of any 2–3 from each section
  - I will have a few reading questions online, too, to get you thinking before class
  - Not everyone will have read every paper, so be ready to:
    - Summarize the author's contribution (if research) or identified issue (if opinion/news)
    - Identify

Assignment Structure

- **15% — "KWLA" Paper [link]**
  - ~7-page paper in total

K. What you already **k**now that pertains to the subject? (~1 page)

W. What you **w**ant to learn about the topic? (~1 page)

L. What you have **l**earned? (~3 pages)

A. How you will **a**pply what you have learned going forward? (~ 1-2 pages)

Assignment Structure

- **20% — Op-Ed/Letter to the Editor/Blog Post** [link]
  - Get practice communicating our work to a *nonexpert* audience
    - Not necessarily the lay public
    - Also potentially developers, or researchers who aren't also linguists
  - How can we:
    - Give an explanation of our research
    - …that is realistic about the risks and possible benefits
    - …while remaining technically accurate?

Assignment Structure

- **50% — Term Project [link]**
  - Option 1 — Run an off-the-shelf system on new data, with demographic information.
  - Option 2 — Analyze an existing task in terms of VSD
  - Option 3 — Some Other Content
- Content:
  - 6-8 page conference-style paper
  - 1-2 page lay summary and risk/benefit analysis

Course Content

1. Intro — Why are we here?
2. Ethical Behavior — Philosophical Underpinnings
3. Value Sensitive Design
4. Accountability — Institutional & Professional Consequences
5. SciComm — Communicating with the Public

Course Content

6. Social Media & Human Subjects
7. Treating Language Data as Ground Truth
8. Bias: in Data & Design, "De-biasing"
9. Abusive Language
10. Best Practices/Wrapup

F.A.T.E.

Fairness

Accountability

Transparency

Ethics

F.A.T.E.
Fairness

- What does being "fair" mean in NLP/Machine Learning?
  - **Process Fairness** vs. **Outcome Fairness**
  - **Individual Fairness** vs. **Group Fairness**

F.A.T.E.
Accountability

- What are the mechanisms by which we correct bad behavior?
  - In our models?
  - In our professional societies?

F.A.T.E.

Transparency

- Do we have good explanations for why we do what we do?
- Why our models do what they do?
- How do we communicate this to the public?

F.A.T.E.
Ethics

- More generally,

Additionally, As Linguists…

- Language is a **proxy** for human intent/behavior [Hovy & Spruit, 2016]
  - …and it is socially conditioned; performative!
  - "*situated*" in language of H&S
- Language is a **proxy** for the state of the world, *as filtered through humans*

Additionally, As Linguists…

- Language is a **proxy** for human intent/behavior [Hovy & Spruit, 2016]
  - …and it is socially conditioned; performative!
  - "*situated*" in language of H&S
- Language is a **proxy** for the state of the world, *as filtered through humans*

As "Data Scientists"

- Statistical modeling is based upon the premise:
  - "*mathematical models, by their nature, are based on the past, and on the assumption that patterns will repeat*" — (O'Neil, 2016)
- Our models only have value insofar as…
  - The past data resembles the present even remotely
  - This assumption is true

Reading