

Natural Language Processing

Lecture 21: Ethics & Safety for NLP

*Many slides & instruction ideas borrowed from:
Mohit Iyyer, David Bamman & Greg Durret*

Expanding Scope of NLP Models

- The impact of NLP is growing universally
 - People interact with NLP systems extensively
 - Conversational agents
 - People make decisions based on NLP systems
 - Ads, recommender systems, ...
 - Legal decisions are being impacted by NLP systems.
 - Paroles, immigration, etc.

Expanding Scope of NLP Models

- People Interact with NLP models throughout their days
 - Conversational agents
- People make decisions based on NLP systems
 - Ads, recommendation systems
- NLP is impacting legal and policy decisions
 - Paroles, immigration decisions, surveillance.

Expanding Scope of NLP Models

In 2020, Uma Mirkhail got a firsthand demonstration of how damaging a bad translation can be.

A crisis translator specializing in Afghan languages, Mirkhail was working with a Pashto-speaking refugee who had fled Afghanistan. A U.S. court had denied the refugee's asylum bid because her written application didn't match the story told in the initial interviews.

In the interviews, the refugee had first maintained that she'd made it through one particular event alone, but the written statement seemed to reference other people with her at the time — a discrepancy large enough for a judge to reject her asylum claim.

After Mirkhail went over the documents, she saw what had gone wrong: An automated translation tool had swapped the "I" pronouns in the woman's statement to "we."

Expanding Scope of NLP Models

The screenshot shows a news article from The Verge. At the top is a black navigation bar with 'THE VERGE' in white, followed by categories: TECH, SCIENCE, CULTURE, CARS, REVIEWS, LONGFORM, VIDEO, and MORE, each with a small dropdown arrow. To the right are social media icons for Facebook, Twitter, and RSS, along with user and search icons.

Below the navigation bar, the breadcrumb trail shows 'US & WORLD \ TECH \ POLITICS \'. The main title of the article is 'Facebook apologizes after wrong translation sees Palestinian man arrested for posting 'good morning''. A small icon with '14' and a speech bubble is to the right of the title. Below the title is a subtitle: 'Facebook translated his post as 'attack them' and 'hurt them''. The author information at the bottom left reads 'by Thuy Ong | @ThuyOng | Oct 24, 2017, 10:43am EDT'.

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

Expanding Scope of NLP Models



[Home](#) [News](#) [Sport](#) [Business](#) [Innovation](#) [Culture](#) [Travel](#) [Earth](#) | [Video](#) [Live](#)

Alexa tells 10-year-old girl to touch live plug with penny

28 December 2021

 Share

Ethics and Safety for NLP

- WWII  Regulation in experimentations with human subjects
 - IRB
 - For NLP/AI?
 -

Ethics & Safety for NLP

- Amplifying the existing Bias
- Exclusion of the underprivileged
- Risks in automation
- Unethical use: harmful usage of systems

Ethics & Safety for NLP

- **Amplifying the existing Bias**
- Exclusion of the underprivileged
- Risks in automation
- Unethical use: harmful usage of systems

Word Embeddings

- Mikolov et al. 2013 show that vector representations have some potential for analogical reasoning through vector arithmetic.

apple - apples \approx car - cars

king - man + woman \approx queen

Science

AAAS

Home

News

Journals

Topics

Careers

Science

Science Advances

Science Immunology

Science Robotics

Science Signaling

Science Translational Medicine

SHARE

REPORT



0



13

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan^{1,*}, Joanna J. Bryson^{1,2,*}, Arvind Narayanan^{1,*}

+ See all authors and affiliations

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230



Peer Reviewed
← see details

Article

Figures & Data

Info & Metrics

eLetters

PDF

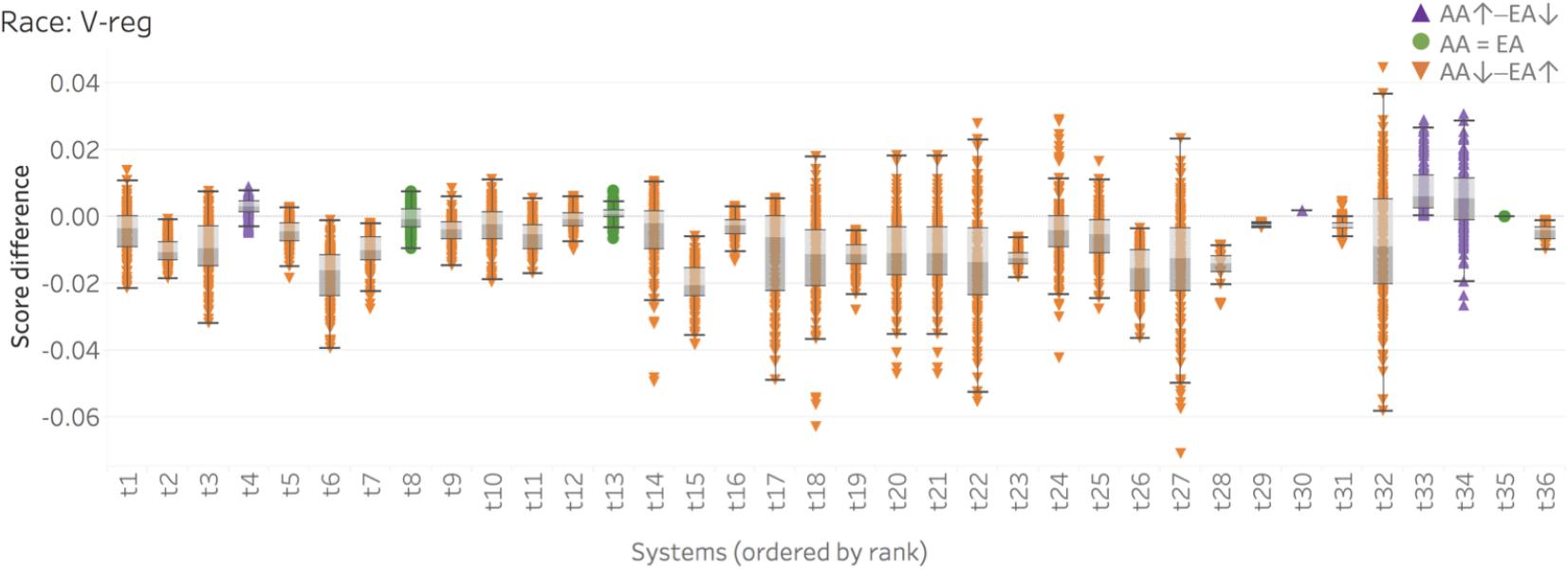
Bias

- Allocational harms: automated systems allocate resources unfairly to different groups (access to housing, credit, parole).
- Representational harms: automated systems represent one group less favorably than another (including demeaning them or erasing their existence).

Representations

- **Pleasant:** caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
- **Unpleasant:** abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit.
- Embeddings for African-American first names are closer to “unpleasant” words than European names (Caliskan et al. 2017)

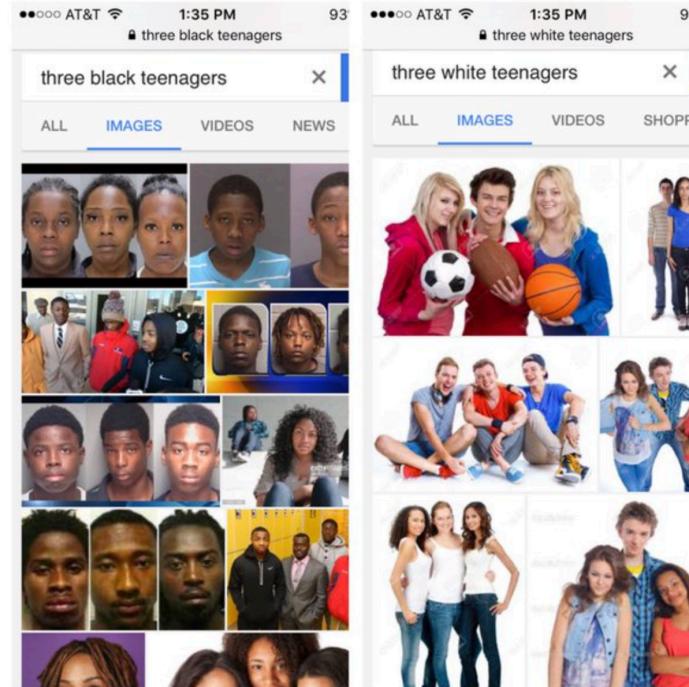
Race: V-reg



- Sentiment analysis over sentences containing African-American first names are more negative than identical sentences with European names

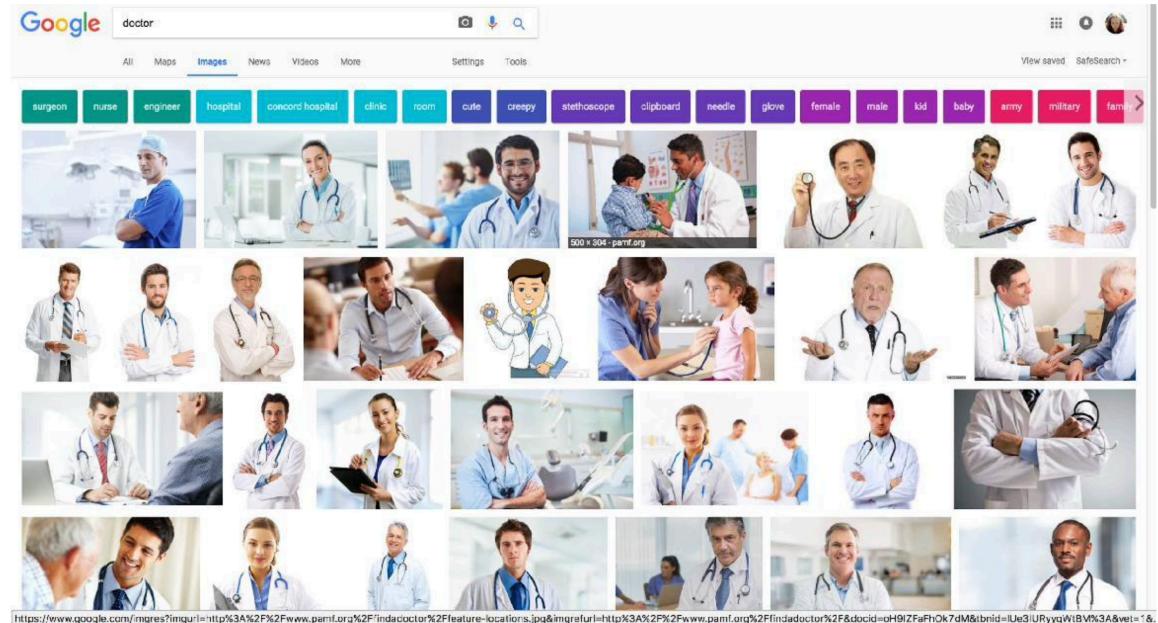
Amplifying the Bias: Race

- In June 2017
 - Three white teenagers
 - Three black teenagers



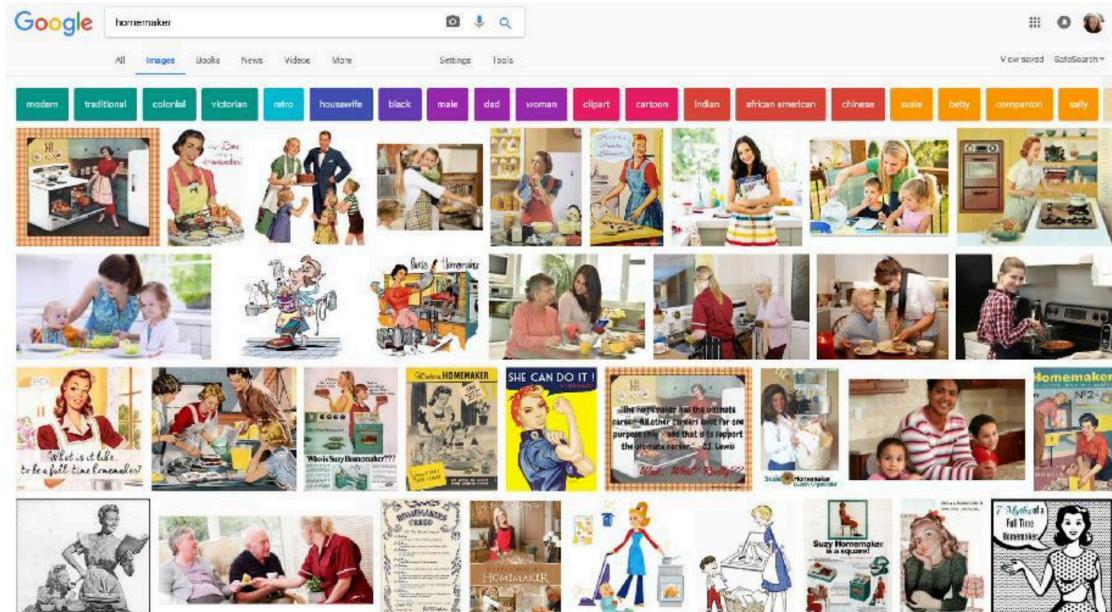
Amplifying the Bias: Gender

- In June 2017



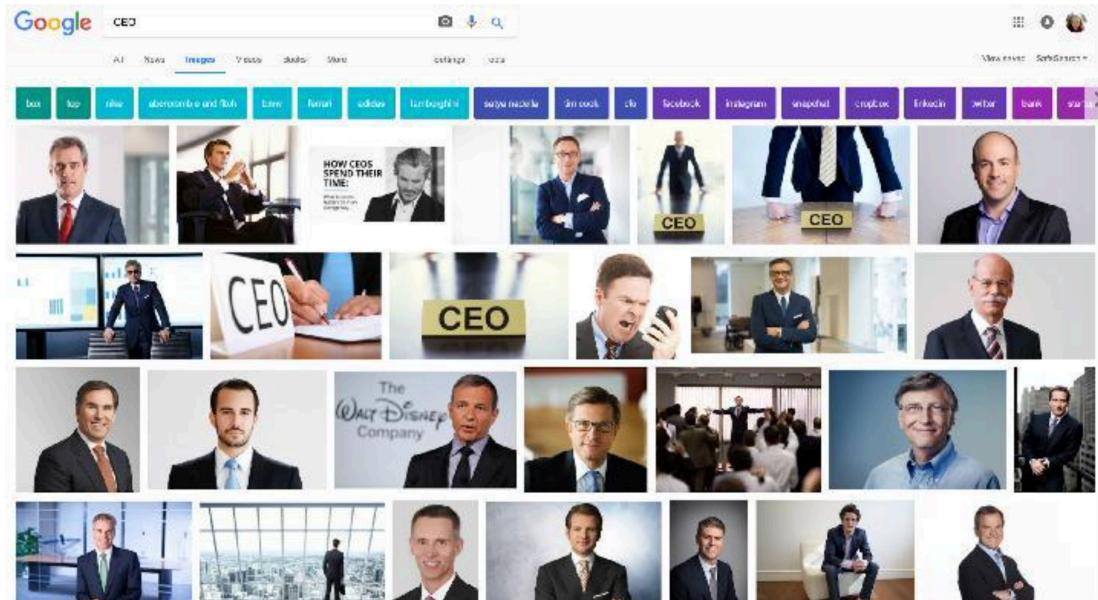
Amplifying the Bias: Gender

- June 2017: image search query “**Homemaker**”



Amplifying the Bias: Gender

- June 2017: image search query “CEO”



Amplifying the Bias: Gender

Mention coref Mention coref Mention coref Mention
The surgeon could n't operate on his patient : it was his son !

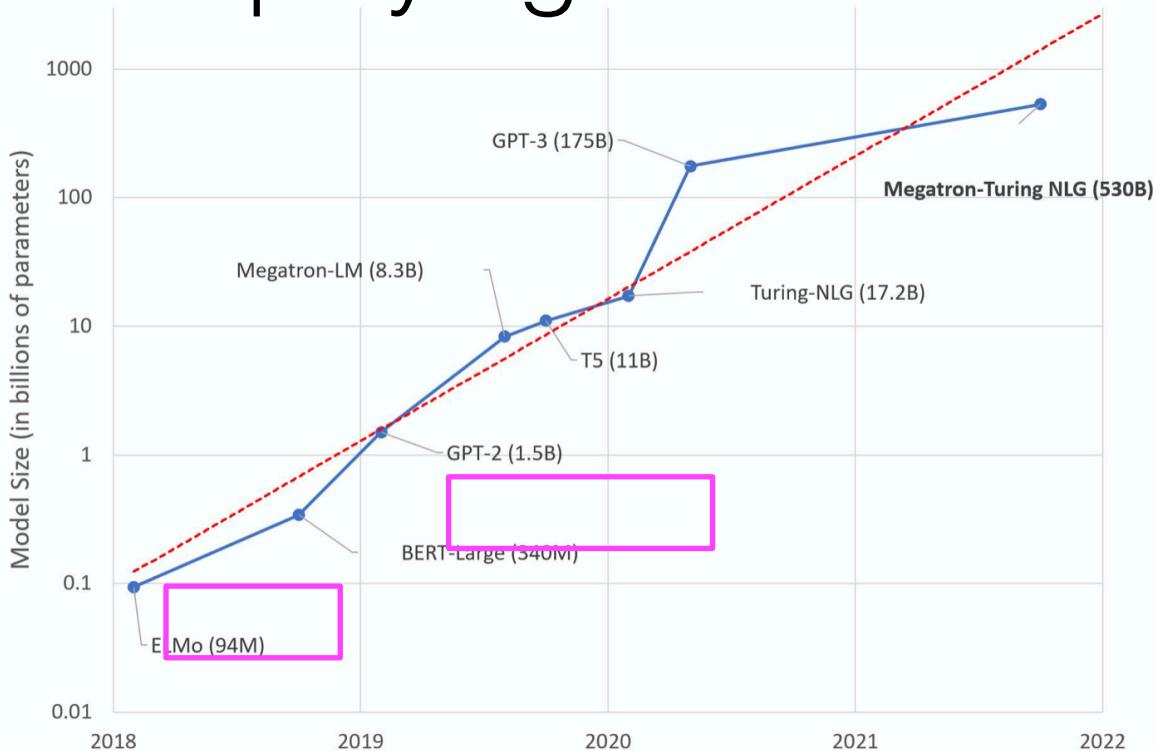
Mention coref Mention coref Mention coref Mention
The surgeon could n't operate on their patient : it was their son !

Mention coref Mention coref Mention coref
The surgeon could n't operate on her patient : it was her son !

Rudinger et al 2018:

Amplifying the Bias: LLMs

Amplifying the Bias: LLMs



<https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/>

Amplifying the Bias: LLMs

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*

ebender@uw.edu

University of Washington
Seattle, WA, USA

Timnit Gebru*

timmnit@blackinai.org

Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major

aymm@uw.edu

University of Washington
Seattle, WA, USA

Shmargaret Shmitchell

shmargaret.shmitchell@gmail.com

The Aether

- Bias in the data: model “size does not guarantee diversity”
- Massive data is challenging to audit, contains data that is biased and is mostly a snapshot of a single point in time.

Amplifying the Bias: LLMs

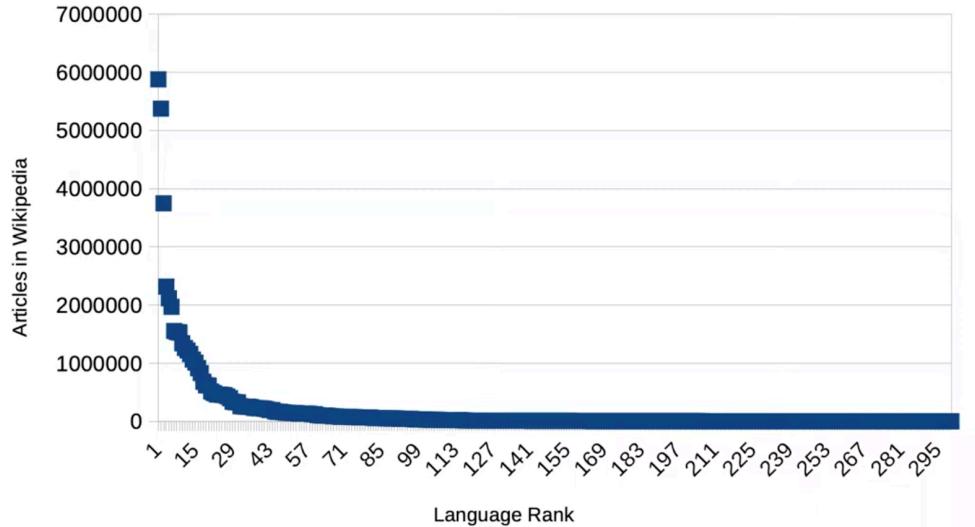
- RLHF reduces to some extent.

Ethics & Safety for NLP

- Amplifying the existing Bias
- **Exclusion of the underprivileged**
- Risks in automation
- Unethical use: harmful usage of systems

Exclusion of the underprivileged

- NLP (and internet content) is predominantly English-centric



Exclusion of the underprivileged

- Presence of people outside the main-stream:
 - Dialects
 - Minorities
 - Elderly
- Treatment of data annotators & turkers
- The burden of cost

Exclusion of the underprivileged

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender*

ebender@uw.edu

University of Washington
Seattle, WA, USA

Timnit Gebru*

timnit@blackinai.org

Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major

aymm@uw.edu

University of Washington
Seattle, WA, USA

Shmargaret Shmitchell

shmargaret.shmitchell@gmail.com

The Aether

- Environmental cost: disproportionately on marginalized populations, who aren't even well-served by models.

Ethics & Safety for NLP

- Amplifying the existing Bias
- Exclusion of the underprivileged
- **Risks in automation**
- Unethical use: harmful usage of systems

Risks in Automation

- NLP systems are impacting decision making

The screenshot shows a news article from The Verge. At the top, there's a navigation bar with links for TECH, SCIENCE, CULTURE, CARS, REVIEWS, LONGFORM, VIDEO, and MORE, along with social media icons for Facebook, Twitter, and RSS. Below the navigation is a header with categories: US & WORLD, TECH, and POLITICS. The main title of the article is "Facebook apologizes after wrong translation sees Palestinian man arrested for posting 'good morning'". A subtext below the title reads "Facebook translated his post as 'attack them' and 'hurt them'". At the bottom of the snippet, it says "by Thuy Ong | @ThuyOng | Oct 24, 2017, 10:43am EDT" and has a view count of 14K.

- Risks of LLM-based annotation
- Risks of LLM-generated WWW.

In 2020, Uma Mirkhail got a firsthand demonstration of how damaging a bad translation can be.

A crisis translator specializing in Afghan languages, Mirkhail was working with a Pashto-speaking refugee who had fled Afghanistan. A U.S. court had denied the refugee's asylum bid because her written application didn't match the story told in the initial interviews.

In the interviews, the refugee had first maintained that she'd made it through one particular event alone, but the written statement seemed to reference other people with her at the time — a discrepancy large enough for a judge to reject her asylum claim.

After Mirkhail went over the documents, she saw what had gone wrong: An automated translation tool had swapped the "I" pronouns in the woman's statement to "we."

Risks in Automation

- The impact on the labor market
-

Risks in Automation

- The impact on the labor market
-

Ethics & Safety for NLP

- Amplifying the existing Bias
- Exclusion of the underprivileged
- Risks in automation
- **Unethical use: harmful usage of systems**

Unethical use of NLP

- Surveillance Systems
- Authorship attribution and de-anonymization

Informante : Paciente varón de 70 años de edad ,
Tagger: PHI NER
minero . Alergias medicamentosas conocidas . Operado de
una hernia el 12 de enero de 2016 en el Hospital Costa del
Sol por la Dra . Juana López . Derivado a este centro el día 16 del
mismo mes para revisión .

Informe clínico del paciente : Paciente SEX de AGE AGE de edad ,
PROFESSION jubilado , sin alergias medicamentosas conocidas .
Operado de una hernia el DATE DATE DATE DATE DATE en el
HOSPITAL HOSPITAL HOSPITAL HOSPITAL por la Dra .
DOCTOR DOCTOR . Derivado a este centro el día 16 del mismo mes
para revisión .

Unethical use of NLP

- Surveillance Systems
- Authorship attribution and de-anonymization

Informe clínico del paciente : Paciente varón de 71 años de edad , biofísico jubilado , sin alergias medicamentosas conocidas . Operado de una hernia el 9 de diciembre de 2021 en el Hospital Alto Jardín * por la Dra . Catalina Reyes . Derivado a este centro el día 16 del mismo mes para revisión .

Informe clínico del paciente : Paciente varón de 69 años de edad , atleta jubilado , sin alergias medicamentosas conocidas . Operado de una hernia el 11 de julio de 2025 en el Hospital Virgen del Palomar por la Dra . Encarnacion Lopez . Derivado a este centro el día 16 del mismo mes para revisión .

Ethics in NLP



- AI2's Delphi:

* Input a **situation** for Delphi to ponder:

Mowing the lawn when there's no grass.

Ponder

Delphi speculates:



Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

“Mowing the lawn when there’s no grass.”

- **You shouldn't**

v1.0.4

<https://delphi.allenai.org/>

Ethics & Safety for NLP

- Amplifying the existing Bias
- Exclusion of the underprivileged
- Risks in automation
- Unethical use: harmful usage of systems

Moving Forward