

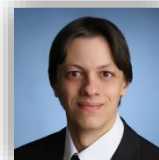
Ethics in Natural Language Processing – SS 2022



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Lecture 3 Bias I

Dr. Thomas Arnold
Aniket Pramanik



Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

Slides and material from Yulia Tsvetkov



Carnegie Mellon University
Language Technologies Institute

Syllabus (tentative)

<u>Nr.</u>	<u>Lecture</u>
01	Introduction, Foundations I
02	Foundations II
03	Bias I
04	Bias II
05	Incivility and Hate Speech I
06	NO LECTURE – Christi Himmelfahrt
07	Incivility and Hate Speech II
08	Low-Resource NLP, NLP for Social Good
09	NO LECTURE - Fronleichnam
10	Privacy and Security I
11	Privacy and Security II
12	Language of Manipulation I
13	Language of Manipulation II

Learning Goals - Example Questions

Describe Kahneman's model of the two cognitive modes ("systems") of our brain.

What is Implicit Bias? How can you recognize / measure it?

Describe the experimental setup of the IAT (Implicit Association Test).

Give two examples of gender bias in web based machine learning systems.

Outline

Implicit Bias

Human Biases in ML

An exercise

Which word is more likely to be used by a **female** ?

Word 1 – Word 2



Which word is more likely to be used by a **female** ?

Giggle – Laugh

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a **female** ?

Giggle – Laugh

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a **female** ?

Brutal – Fierce

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a **female** ?

Brutal – Fierce

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a **older person** ?

Impressive – Amazing

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a **older person** ?

Impressive – Amazing

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a person of **higher occupational class** ?

Suggestions – Proposals

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a person of **higher occupational class** ?

Suggestions – **Proposals**

(Preotiuc-Pietro et al. '16)

Why do we intuitively recognize a default social group?

Implicit Bias

How Do We Make Decisions

System 1

automatic

fast

parallel

automatic

effortless

associative

slow-learning

System 2

effortful

slow

serial

controlled

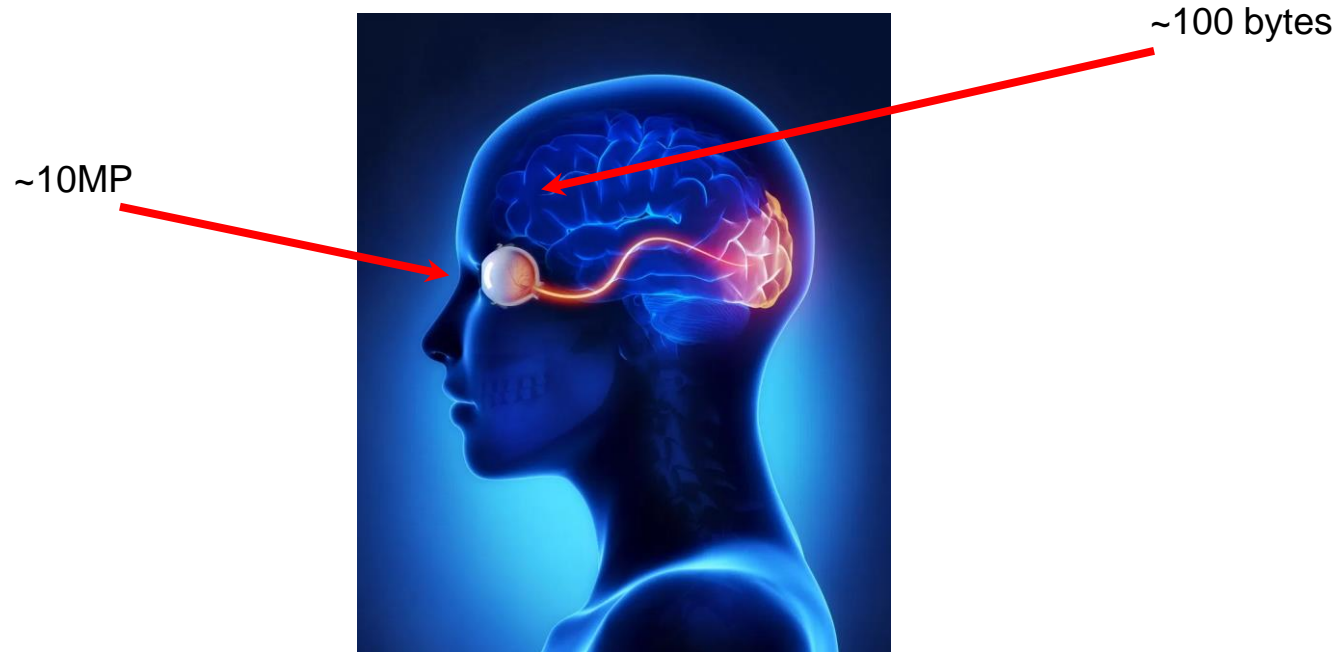
effort-filled

rule-governed

flexible

Kahneman & Tversky 1973, 1974, 2002

Why?



How Do We Make Decisions

System 1

automatic

System 2

effortful

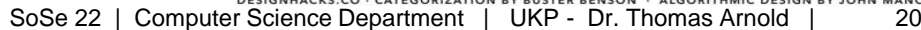
Our brains are evolutionarily hard-wired to store learned information for rapid retrieval and automatic judgments. Over 95% of cognition is relegated to the System 1 “auto-pilot.”

Psychological Perspective on Implicit Bias

Biases inevitably form because of our innate tendency to:

- **Categorize** the world to simplify processing
- **Store** learned information in mental representations (called schemas)
- Automatically and unconsciously **activate** stored information whenever one encounters a category member

Cognitive bias is a systematic pattern of deviation from rationality in judgement



Biases affect how we make decisions

- **confirmation bias**: paying more attention to information that reinforces previously held beliefs and ignoring evidence to the contrary
- **ingroup favoritism**: when one favors in-group members over out-group members
- **group attribution error**: when one generalizes about a group based on a group of representatives
- **halo effect**: when overall impression of a person impacts evaluation of their specific traits
- etc.

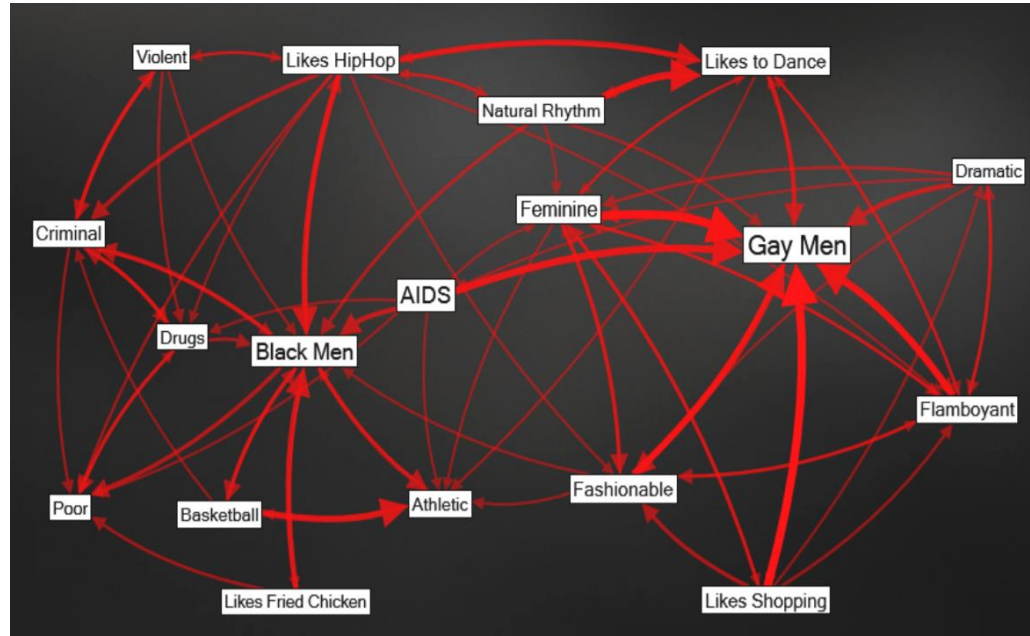




Social stereotypes

- Gender
- Race
- Disability
- Age
- Sexual orientation
- Culture
- Class
- Poverty
- Language
- Religion
- National origin
- ...



Social stereotypes are similarly internalized as associations through natural processes of learning and categorization



Implicit biases are pervasive, unconscious, and can automatically influence the ways in which we see and treat others, even when we are determined to be fair and objective.

How to Recognize Implicit Bias?

Implicit Association Test - Greenwald et al. 1998

Category	Items
Good	Spectacular, Appealing, Love, Triumph, Joyous, Fabulous, Excitement, Excellent
Bad	Angry, Disgust, Rotten, Selfish, Abuse, Dirty, Hatred, Ugly
African Americans	
European Americans	

Implicit Association Test

GOOD

BAD

Love

Implicit Association Test

GOOD

BAD

Hatred

Implicit Association Test

GOOD

BAD

Spectacular

Implicit Association Test

**African
Americans**

**European
Americans**



Implicit Association Test

**African
Americans**

**European
Americans**



Implicit Association Test

**African
Americans**

**or
GOOD**

**European
Americans**

**or
BAD**

Appealing

Implicit Association Test

**African
Americans**

**or
GOOD**

**European
Americans**

**or
BAD**



Implicit Association Test

**African
Americans**

or

GOOD

**European
Americans**

or

BAD



Implicit Association Test

**African
Americans**

**or
GOOD**

**European
Americans**

**or
BAD**

Rotten

Implicit Association Test

BAD

GOOD

Love

Implicit Association Test

**African
Americans**

**or
BAD**

**European
Americans**

**or
GOOD**

Spectacular

Implicit Association Test

**African
Americans**

**or
BAD**

**European
Americans**

**or
GOOD**



Implicit Association Test

**African
Americans**

**or
BAD**

**European
Americans**

**or
GOOD**



Implicit Association Test

The IAT involves making repeated judgments (by pressing a key on a keyboard) to label words or images that pertain to one of two categories presented simultaneously (e.g., categorizing pictures of African American or European American and categorizing positive/negative adjectives).

The test compares response times when different pairs of categories share a **response key** on keyboard

(e.g., African American + GOOD vs African American + BAD vs European American + GOOD vs European American + BAD)

How Implicit Bias Manifests?

Micro-inequities

Micro-inequities: ephemeral, covert, unintentional, frequently unrecognized events
that reinforce power dynamics or perceptions of “difference”

slights, exclusions, slips of the tongue, nonverbal signals, unchecked assumptions, unequal expectations, etc.



Slide credit: Geoff Kaufman

Microaggressions

“A comment or action that **subtly and often unconsciously or unintentionally** expresses a prejudiced attitude towards a member of a marginalized group”

- Merriam Webster

Surface-level sentiment can be negative, neutral, or positive. For example:

- “Girls just **aren’t good** at math.”

microaggressions.com
tumblr.

Microaggressions

“A comment or action that **subtly and often unconsciously or unintentionally** expresses a prejudiced attitude towards a member of a marginalized group”

- Merriam Webster

Surface-level sentiment can be negative, neutral, or positive. For example:

- “Girls just **aren’t good** at math.”
- “Don’t you people like tamales?”

microaggressions.com
tumblr.

Microaggressions

“A comment or action that **subtly and often unconsciously or unintentionally** expresses a prejudiced attitude towards a member of a marginalized group”

- Merriam Webster

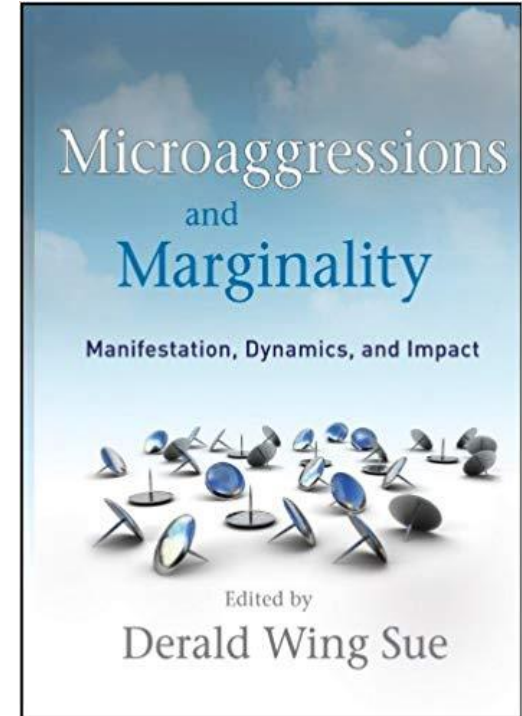
Surface-level sentiment can be negative, neutral, or positive. For example:

- “Girls just **aren’t good** at math.”
- “Don’t you people like tamales?”
- You’re **too pretty** to be gay.”

microaggressions.com
tumblr.

Harmful impact of microaggressions

- Effects can be more pernicious (~harmful) than overtly aggressive speech (Sue et al. 2007, Sue 2010, Nadal et al. 2014)
- Can affect people's professional experiences and career trajectories (Cortina et al. 2002, Trix and Psenka 2003)
- Play on, and reinforce, problematic stereotypes and power structures (Hall and Braunwald 1981, Fournier et al. 2002)



Stereotype threat

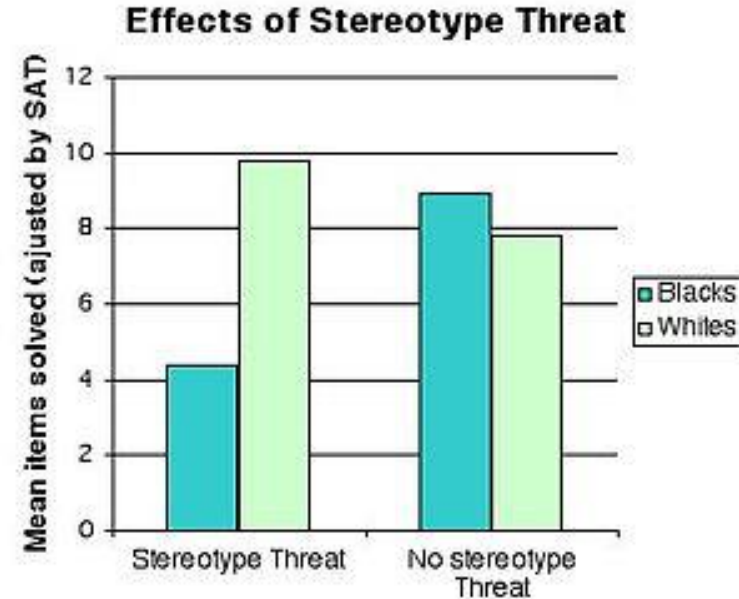
Fear of confirming a negative stereotype about one's group (Steele & Aronson, 1995)

- Often leads to anxiety and negative feelings that can use up mental resources and undermine one's confidence and ability to succeed
- Exacerbated by repeated experiences with microaggressions reducing one's sense of belonging or self-belief in a particular domain
 - e.g., women in STEM: Beasley & Fischer'12; Shapiro & Williams'12

Stereotype threat

- Groups: Blacks and Whites
- Threat: Intellectual ability

J. Aronson, C.M. Steele, M.F. Salinas, M.J. Lustina,
Readings About the Social Animal, 8th edition, ed. E. Aronson



Back to AI

Back to AI

AI is only "System 1"

Terminology

Cognitive bias



Statistical bias in ML



Social biases in AI, in data, algorithms, and applications

Outline

Implicit Bias

Human Biases in ML



- Conversational agents
- Personal assistants
- Search engines
- Translation engines
- Medical research assistants

Gender Bias on the Web

- The dominant class is often portrayed and perceived as relatively more professional (Kay, Matuszek, and Munson 2015)
- Males are over-represented in the reporting of web-based news articles (Jia, Lansdall-Welfare, and Cristianini 2015)
- Males are over-represented in twitter conversations (Garcia, Weber, and Garimella 2014)
- Biographical articles about women on Wikipedia disproportionately discuss romantic relationships or family-related issues (Wagner et al. 2015)
- IMDB reviews written by women are perceived as less useful (Otterbacher 2013)



Online data is riddled with **SOCIAL STEREOTYPES**



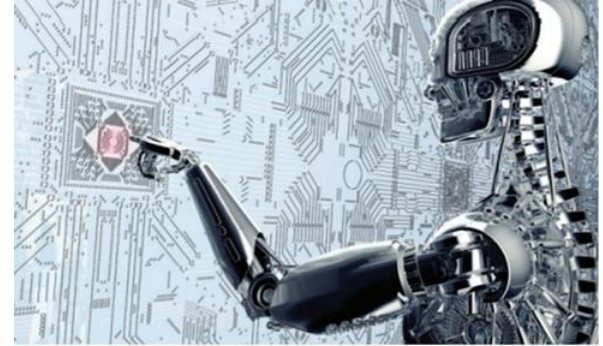
Online data is riddled with **SOCIAL STEREOTYPES**



Consequence: models are also biased

Usage of AI Systems

- employment matching
- flight routing
- automated legal aid in immigration
- advertisement placement
- search
- parole decisions
- chatbots
- face recognition
- voice recognition
- + dozens!



An Intelligence in Our Image

The Risks of Bias and Errors in
Artificial Intelligence

Osonde Osoba, William Welser IV



Discussion

User-generated content represents “real world data”.
Is it wrong to build models replicating real world data?

Image Search

- June 2016: web search query “three black teenagers”

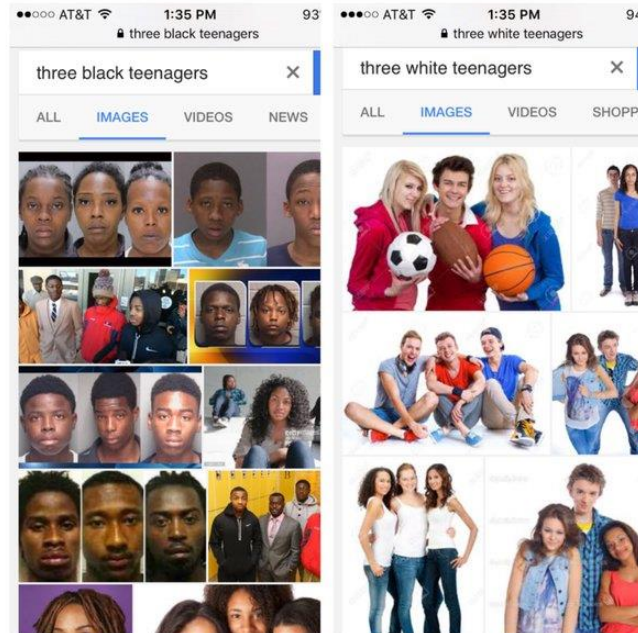


Image Search

- June 2017: image search query “Doctor”

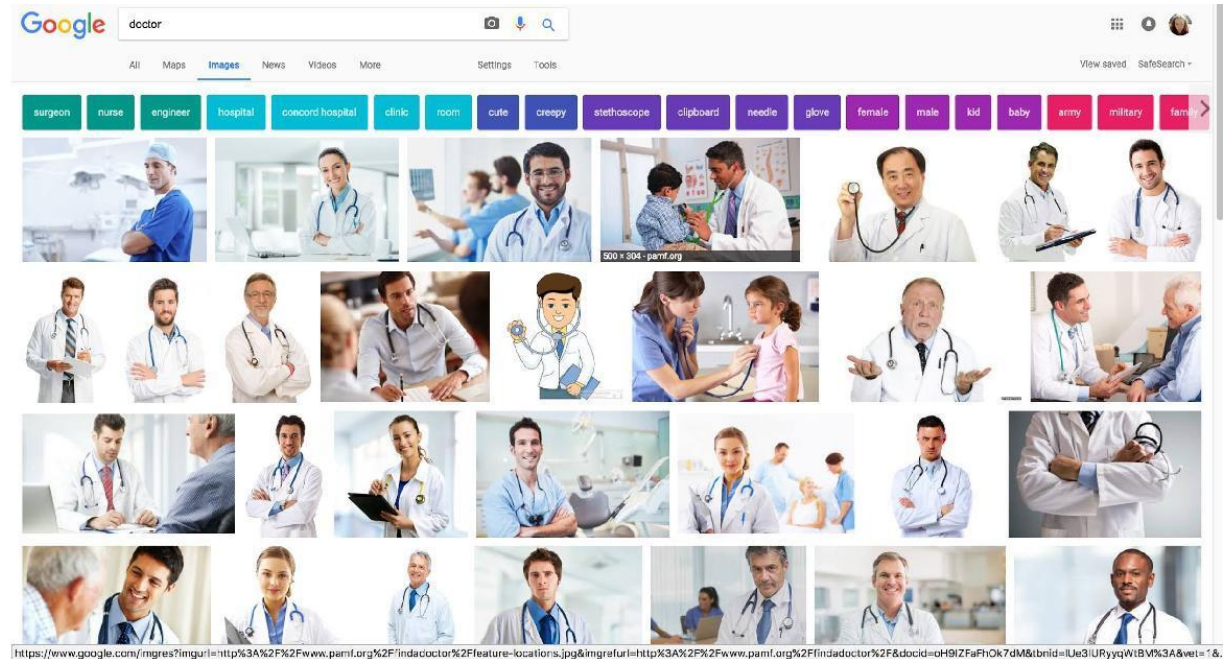


Image Search

- June 2017: image search query “Nurse”

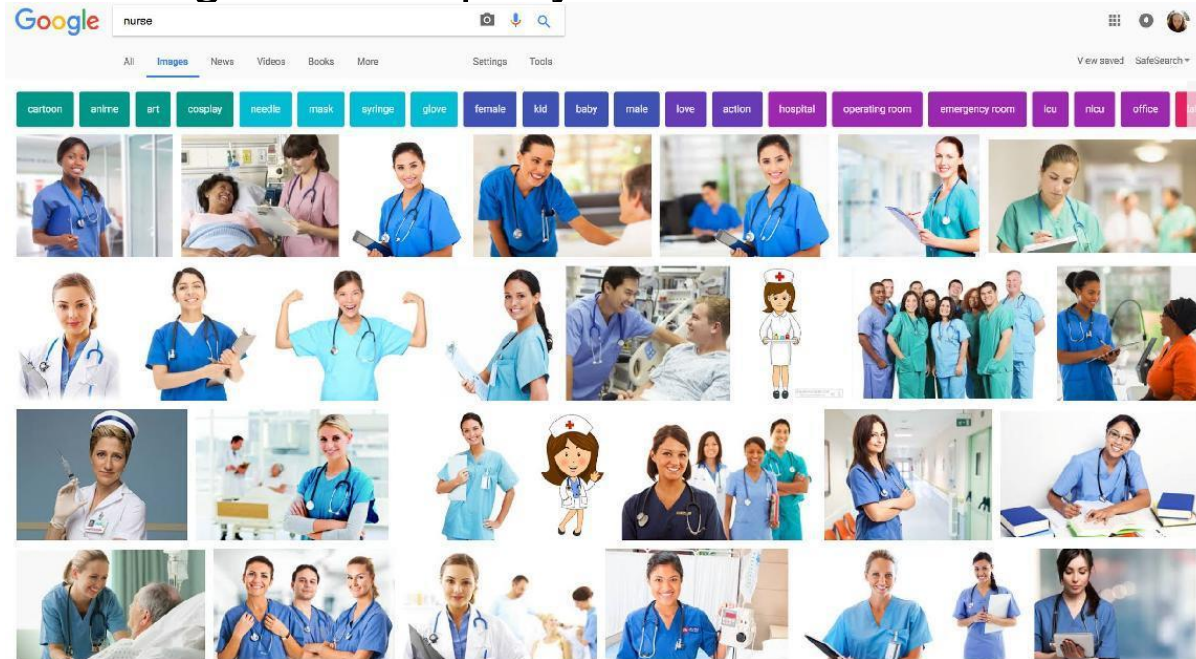


Image Search

- June 2017: image search query “Homemaker”

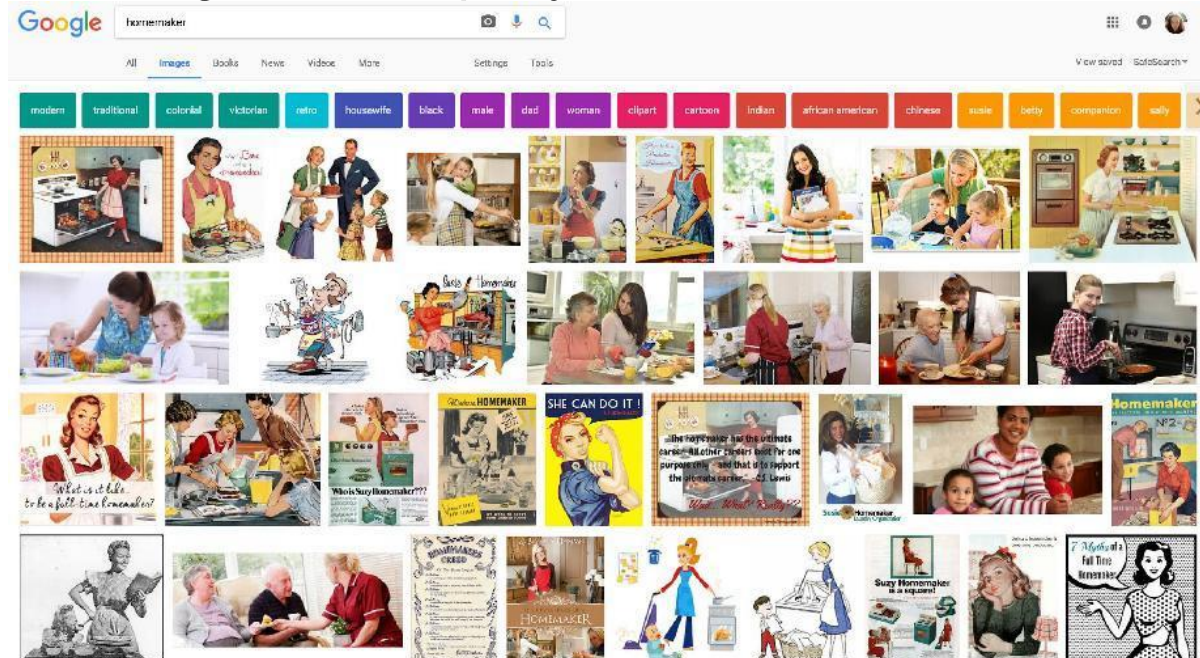
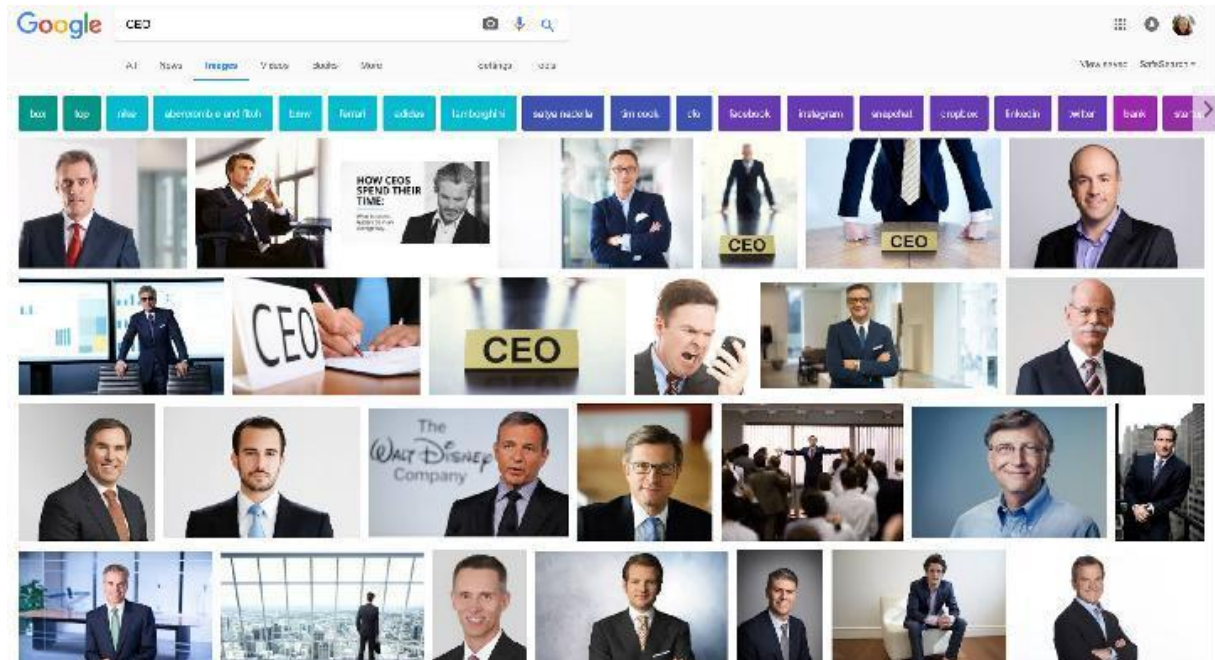


Image Search

- June 2017: image search query “CEO”



- Google professor

Images News Videos Books More Settings Tools

View saved SafeSearch

hot female android male baby african american indian chinese japanese university college classroom lab concord hospital cartoon meme comic science fashion

Image Search

- May 2022: image search query “CEO” = not much better...

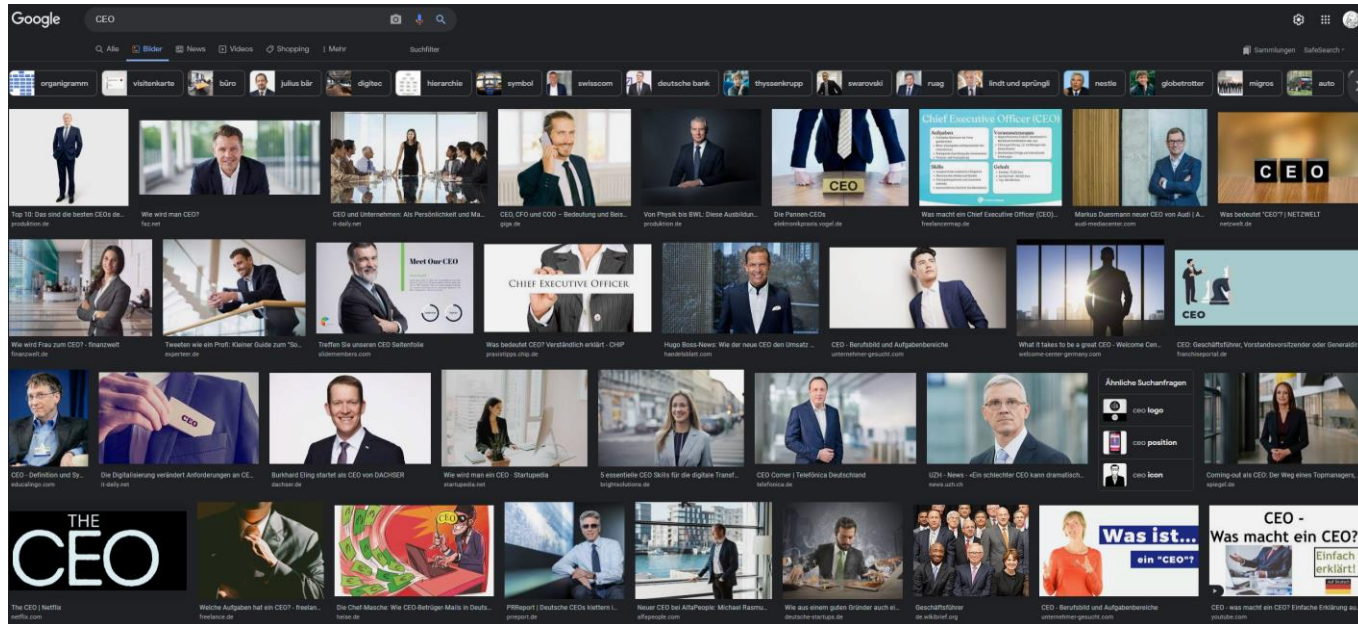
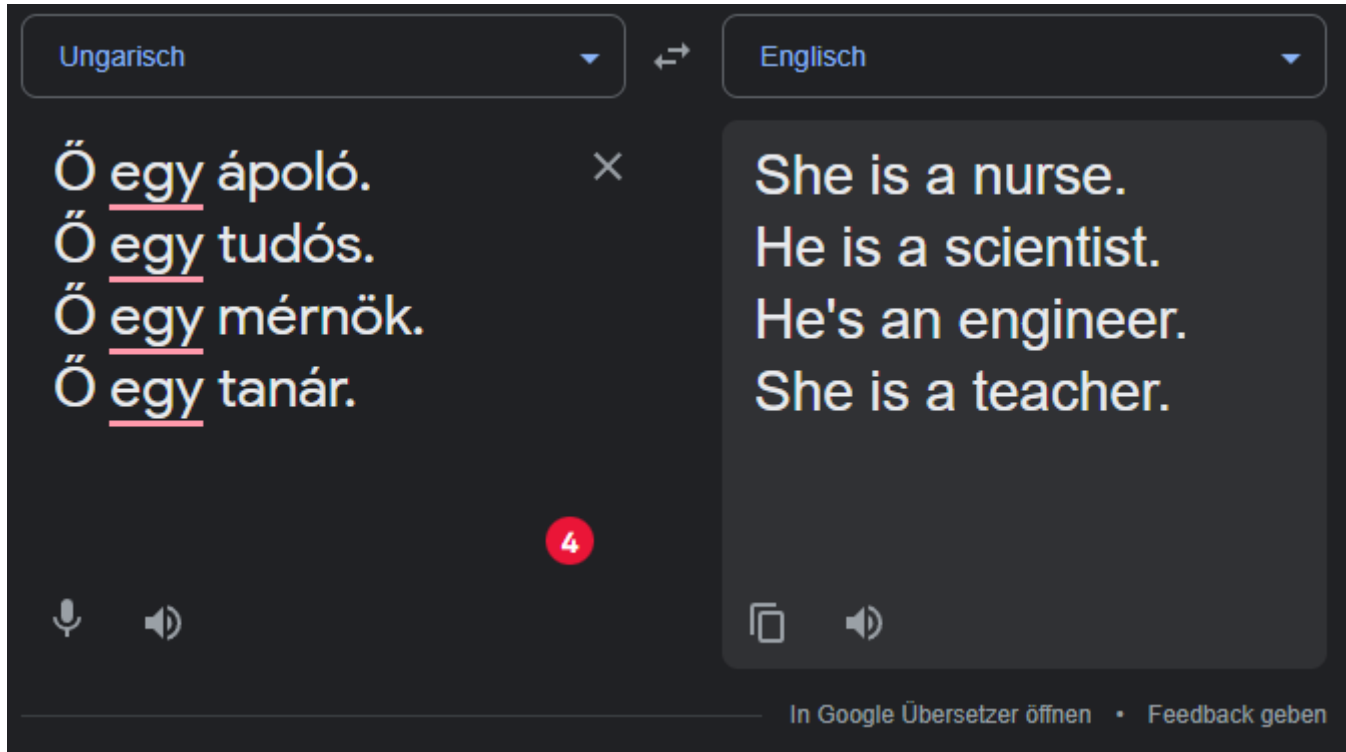


Image Search

- May 2022: image search query “Doctor” = a bit better...



Bias in Machine Translation



The screenshot shows the Google Translate interface with 'Ungarisch' (Hungarian) on the left and 'Englisch' (English) on the right. The Hungarian input text is:

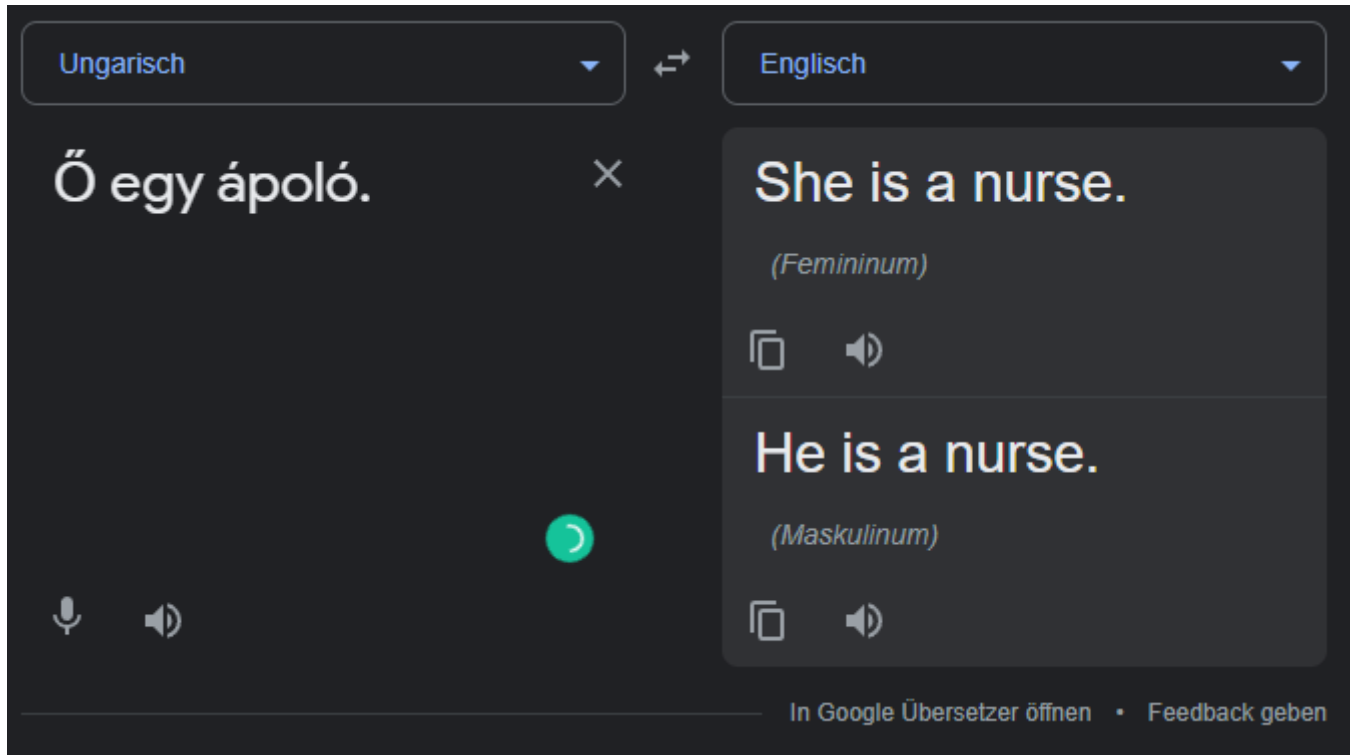
Ő egy ápoló.
Ő egy tudós.
Ő egy mérnök.
Ő egy tanár.

The English output text is:

She is a nurse.
He is a scientist.
He's an engineer.
She is a teacher.

The word 'egy' is underlined in the Hungarian text, and the corresponding English translations are highlighted in yellow. A red circle with the number '4' is visible in the bottom left corner of the interface. At the bottom, there are links for 'In Google Übersetzer öffnen' and 'Feedback geben'.

Bias in Machine Translation



Applications

- Machine Translation
- Caption generation
- Speech Recognition
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

Core technologies

- Language modelling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense
disambiguation
- Semantic Role Labelling
- ...

Applications

- Machine translation (Douglas'17, Prates et al. '19)
- Caption generation (Burns et al.'18)
- Speech recognition (Tatman'17)
- Question answering (Burghardt et al. '18)
- Dialogue systems (Dinan, Fan et al. '19)
- Summarization (Jung, Kang et al. '19)
- Sentiment analysis (Kiritchenko & Mohammad '18)
- Language identification (Blodgett et al.'16, Jurgens et al.'17)
- Text classification (Dixon et al. '18, Sap et al. '19, Kumar et al. '19)

Core technologies

- Language modeling (Lu et al. '18)
- Named-entity recognition (Mehrabi et al. '19)
- Coreference resolution (Zhao et al. '18, Rudinger et al. '18)
- Semantic role labelling (Zhao et al. '17)
- SNLI (Rudinger et al. '17)
- Word embeddings (Bolukbasi et al. '16, Caliskan et al.'17,++)
- ...
- **Surveys** (Sun&Gaut et al.'19, Blodgett et al.'20, Field et al.'21)

Next Lecture

- Bias in ML predictions
- Bias Amplification
- Debiasing techniques

Next Lecture

Bias II