

# Ethics in Natural Language Processing

Pierre Lison

**IN4080**: Natural Language  
Processing (Fall 2022)

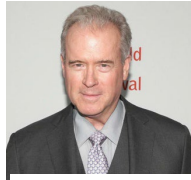
03.11.2022



# Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

**Whistleblower describes how firm linked to former Trump adviser Steve Bannon compiled user data to target American voters**

- **T made Steve Bannon's psychological warfare tool: meet the data war whistleblower**
- **Mark Zuckerberg breaks silence on Cambridge Analytica**



## Opinion **Artificial intelligence**

Trusting AI too much can turn out to be fatal

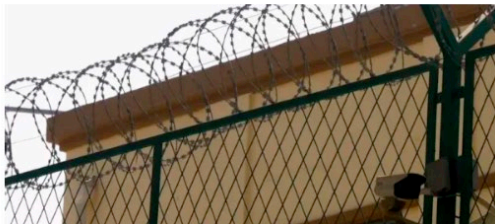
We follow faulty automated instructions because 'the computer can't be wrong'

JOHN THORNHILL

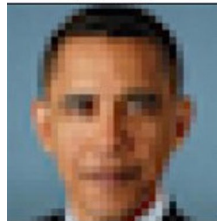
+ Add to myFT



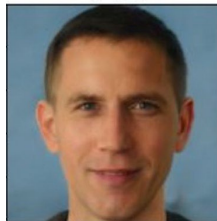
## How China's Government Is Using AI on Its Uighur Muslim Population



Original



Result



## Chatbots and Abuse: A Growing Concern

## How IBM Watson Overpromised and Underdelivered on AI Health Care

After its triumph on *Jeopardy!*, IBM's AI seemed poised to revolutionize medicine. Doctors are still waiting

# Plan for today

- ▶ What is ethics?
- ▶ Misrepresentation & bias
- ▶ Unintended consequences
- ▶ Misuses of technology
- ▶ Privacy & trust

# Plan for today

- ▶ **What is ethics?**
- ▶ Misrepresentation & bias
- ▶ Unintended consequences
- ▶ Misuses of technology
- ▶ Privacy & trust

# Ethics

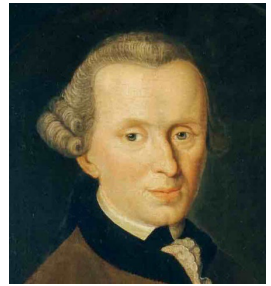
- ▶ = the systematic study of conduct based on moral principles, reflective choices, and standards of **right** and **wrong** conduct

[P. Wheelwright (1959), *A Critical Introduction to Ethics*]

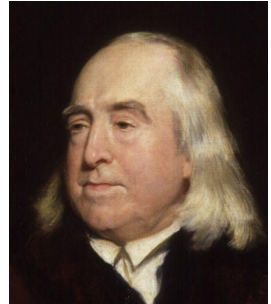
- ▶ *A practical discipline* – how to act?
- ▶ Depends on our values, norms and beliefs
  - No unique, objective answers!
  - But more than just "opinions" – need to *justify* our choices in a rational manner

# Ethics

- ▶ Various philosophical traditions to define what is good/bad:
  - **Deontological:** respect of *moral principles* and rules
  - **Consequentialist:** focus on the *outcomes* of our actions
  - (and more)
- ▶ No particular "side" in this lecture
  - Inspiration from multiple ethical perspectives



Immanuel Kant  
(1724-1804)



Jeremy Bentham  
(1748-1832)

# Ethics



← Protests against LAPD's system for "data-driven predictive policing"

← Protests against Facebook's perceived passivity against disinformation campaigns (fake news etc.)

The NLP tools we build, deploy or maintain have **real impacts** on **real people**

- Who might benefit/be harmed?
- Can our work be misused?
- Which objective do we optimise?

# Ethics

- ▶ Ethical behaviour is a basis for **trust**
- ▶ We have a **professional duty** to consider the ethical consequences of our work
- ▶ Ethical  $\neq$  Legal!
  - Plenty of actions are not illegal but will be seen by most as unethical
  - Laws should embody moral principles - but don't always do





# Plan for today

- ▶ What is ethics?
- ▶ **Misrepresentation & bias**
- ▶ Unintended consequences
- ▶ Misuses of technology
- ▶ Privacy & trust

# Language and people

"The common misconception is that language has to do with words and what they mean.

It doesn't.

It has to do with **people** and what **they** mean."

[H. Clark & M. Schober (1992), "Asking questions and influencing answers", *Questions about questions.*]

Language data does not exist in a vacuum – it comes from *people* and is used to communicate with other people!

- These people may have various *stereotypes & biases*
- & their relative position of *power* and *privilege* affects the status of their language productions

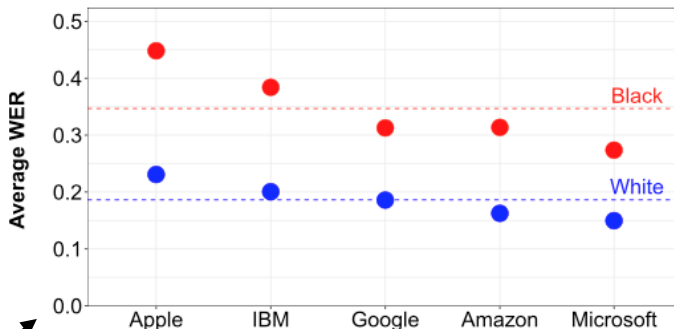
# Demographic biases

- ▶ Certain demographic groups are largely *over-represented* in NLP datasets
  - That is, the proportion of content from these groups is >> their demographic weight
  - Ex: young, educated white males from US
- ▶ Under-representation of linguistic & ethnic minorities, low-educated adults, etc.
  - & gender: 16% of female editors in Wikipedia (and 17% of biographies are about women)



# Demographic biases

- ▶ Under-represented groups in the training set of an NLP model will often experience *lower accuracies* at prediction time



[A. Koenecke et al (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*]

- ▶ Can lead to the *technological exclusion* of already disadvantaged groups

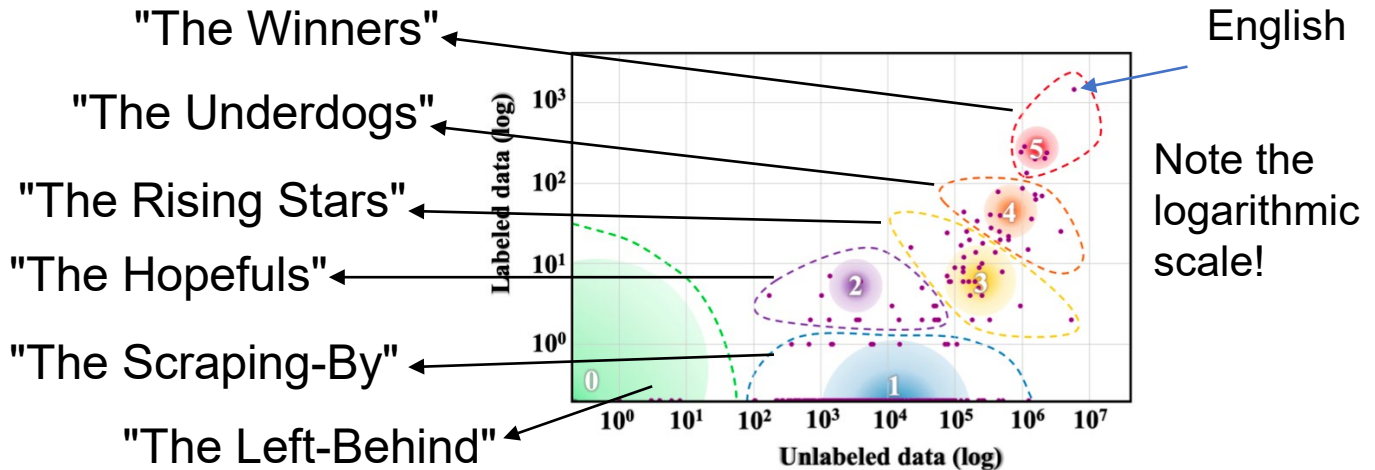
# Elderly users



Sketch from Saturday Night Live, 2017

# Linguistic (in)justice

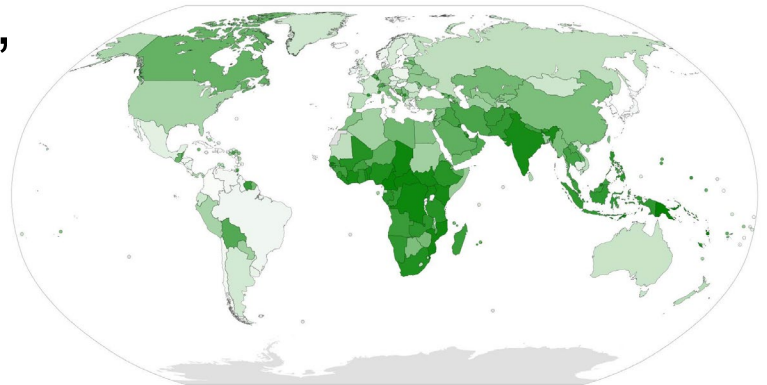
Only a small fraction of the world's 7000 languages covered in NLP datasets & tools



[Joshi et al (2020), The State and Fate of Linguistic Diversity and Inclusion in the NLP World, ACL]

# Linguistic (in)justice

- ▶ The lack of linguistic resources & tools for most languages is a huge ethical issue
- ▶ We exclude from our technology the part of the world's population that is already most vulnerable, both culturally and socio-economically!



Linguistic diversity index

# Linguistic (in)justice

The dominance of US & British English in NLP is also a scientific problem

- ▶ NLP research not sufficiently exposed to *typological variety*
- ▶ Focus on linguistic traits that are important in English (such as word order)
- ▶ Neglect of traits that are absent or minimal in English (such as morphology)





# Social biases

Stereotypes, prejudices, sexism & other types of social biases expressed in the training data will also creep into our NLP models

NORWEGIAN - DETECTED

ENGLISH

SPANISH



GERMAN

ENGLISH

SPANISH



Legen ba sekretæren om hjelp.



Der Arzt bat die Sekretärin um Hilfe.



29/5000



Masculine form

Feminine form



# Social biases

Also observed in language modelling:



And even in  
word  
embeddings:

- Extreme *she* occupations**
- |                 |                       |                        |
|-----------------|-----------------------|------------------------|
| 1. homemaker    | 2. nurse              | 3. receptionist        |
| 4. librarian    | 5. socialite          | 6. hairdresser         |
| 7. nanny        | 8. bookkeeper         | 9. stylist             |
| 10. housekeeper | 11. interior designer | 12. guidance counselor |

- Extreme *he* occupations**
- |                |                   |                |
|----------------|-------------------|----------------|
| 1. maestro     | 2. skipper        | 3. protege     |
| 4. philosopher | 5. captain        | 6. architect   |
| 7. financier   | 8. warrior        | 9. broadcaster |
| 10. magician   | 11. fighter pilot | 12. boss       |



[Bolukbasi, T. et al (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *NIPS*.]

# Social biases

- ▶ NLP models may not only **reflect** but also **amplify** biases in training data
  - & make biases appear more "objective"
- ▶ Harms caused by social biases are often *diffuse, unconscious & non-intentional*
  - More pernicious & difficult to address!
  - Relatively small levels of harm ("microaggressions"), but experienced *repeatedly by whole social groups*



# Debiasing

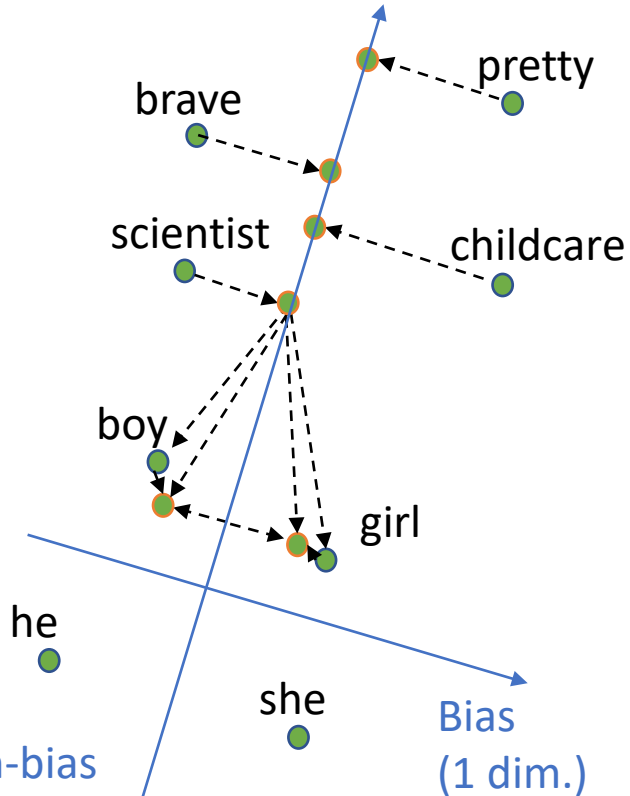
1. Identify bias *direction*  
(more generally: subspace)

$$\left. \begin{array}{l} \overrightarrow{boy} - \overrightarrow{girl} \\ \overrightarrow{he} - \overrightarrow{she} \\ \dots \end{array} \right\} \text{take average}$$

2. "*Neutralise*" words that are not definitional  
(=set to zero in bias direction)

3. Equalise pairs  
(such as "boy" – "girl")

Non-bias  
(299 dim.)



[Bolukbasi, T. et al (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *NIPS*.]

# Gender in MT

- ▶ In languages with grammatical gender, the speaker gender may affect translation:

**English:** [M/F] I'm happy

**French:** Je suis heureux (*if speaker is male*)

Je suis heureuse (*if speaker is female*)

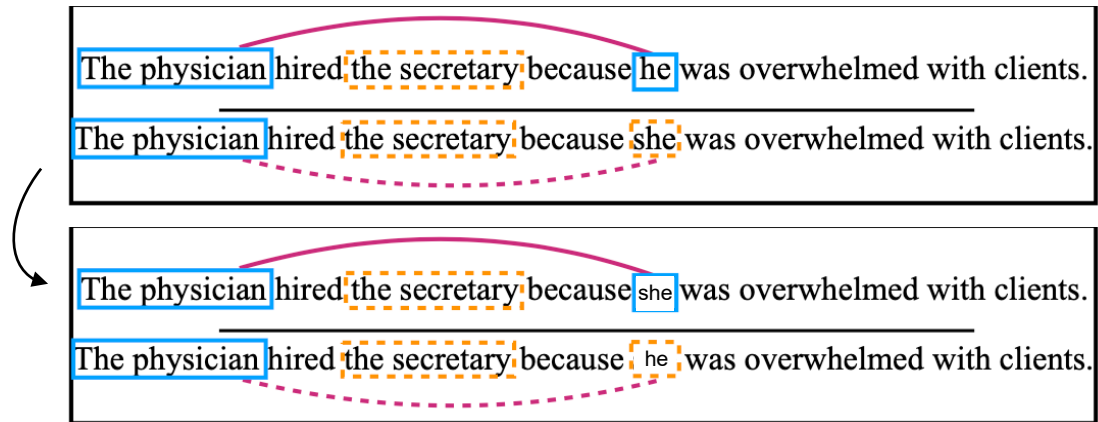
- ▶ Male-produced texts are dominant in translated data → male bias in MT
- ▶ Solution: **tag** the speaker gender



[Vanmassenhove, E., et al (2018). Getting Gender Right in Neural Machine Translation. In EMNLP]

# Debiasing

One easy debiasing method is through *data augmentation*, i.e. by adding gender-swapped examples to the training set



# Social biases

- ▶ Biases can also creep in data annotations (categories, output strings etc.)
- ▶ Annotations are never neutral, they are a *prism* through which we see the world



"loser"



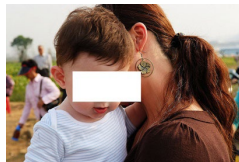
"swinger"



"toy"



"hermaphrodite"



"mixed-blood"

Those are **real labels** in *ImageNet*, the most widely used dataset in computer vision !

[K. Crawford & T. Paglen (2019)  
"Excavating AI: The politics of images  
in machine learning training sets"]

# Fairness

We want our systems to be **fair**. What does that mean?

- ▶ Imagine a group of individuals distinguished by a *sensitive attribute*  $A$ , like race or gender
- ▶ Each individual has a feature vector  $\mathbf{X}$ , and we wish to make a prediction  $\hat{Y}$  based on  $\mathbf{X}$



**Example:** predict the likelihood of *recidivism* among released prisoners, while ensuring our predictions are not racially biased



# Definitions of fairness

1. **Unawareness:** require that the features  $\mathbf{X}$  leave out the sensitive attribute  $A$ 
  - *Problem:* ignores correlations between features (such as the person's neighbourhood)
2. **Demographic parity:**

$$P_{A=1}(\hat{Y} = 1) \approx P_{A=0}(\hat{Y} = 1)$$

In our example, this would mean that the proportion of prisoners predicted to become recidivists should be (approx.) the same for whites and non-whites

# Definitions of fairness

## 3. Predictive parity: (with $y = 0$ and $1$ )

$$P_{A=1}(Y = y | \hat{Y} = y) \approx P_{A=0}(Y = y | \hat{Y} = y)$$

→ The *precision* of our predictions (recidivism or not) should be the same across the two groups

## 4. Equality of odds

$$P_{A=1}(\hat{Y} = y | Y = y) \approx P_{A=0}(\hat{Y} = y | Y = y)$$

→ The *recall* of our predictions should be the same. In particular, if I am not going to relapse to crime, my odds of being marked as recidivist should be similar

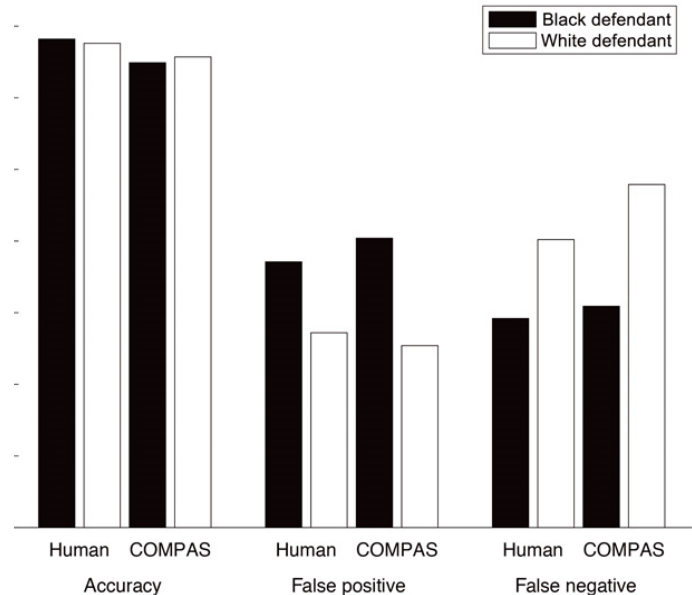
# Fairness

[Friedler, S. A. et al (2016).  
*On the (im)possibility of fairness*]

▶ Those fairness criteria are incompatible - cannot satisfy them simultaneously!

▶ **COMPAS** software:

- Optimised for predictive parity
- Led to biased *odds* (black defendants much more likely to be false positives)



- ▶ What is ethics?
- ▶ Misrepresentation & bias
- ▶ **Unintended consequences**
- ▶ Misuses of technology
- ▶ Privacy & trust

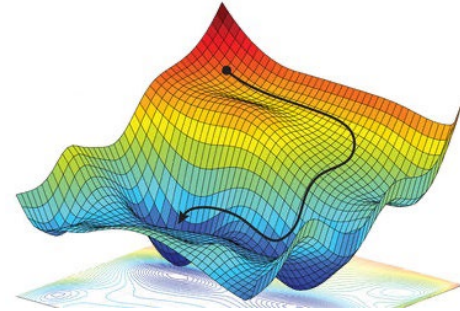
# Unintended consequences

As computing professionals, we have a duty to consider how the software we develop may be used in practice.



What may be the (intended or unintended) **impacts** of this software on individuals, social groups, or society at large?

# Training objectives



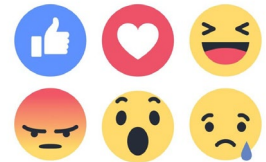
ML models are built to optimise a given **objective function**

- Or, equivalently, minimise a **loss function**
- In classification, we often seek to minimise the *cross-entropy loss* between the model predictions and the actual labels
- In reinforcement learning, we maximise the *expected cumulative reward*

► The objective function defines what we perceive as *good solutions* for a task

# Training objectives

- ▶ "Externalities" that are not part of the objective function are thus ignored
- ▶ *Example:* ML models used in social media are often optimized towards maximising *user engagement*
  - As it turns out, controversial & divisive content yields more user engagement on social media
  - Which leads to wide-ranging consequences, such as heightened *political polarisation*



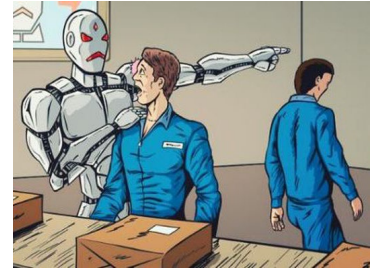
# Training objectives

- ▶ Ideally, we wish our objective function to include all factors that we consider part of a "good solution"
  - Such as not increasing political polarisation
- ▶ Not always possible in practice, specially for factors that cannot be easily *measured*
  - But we must be aware of the discrepancy between what we would consider as "good solutions" and what we actually optimise





# Automation

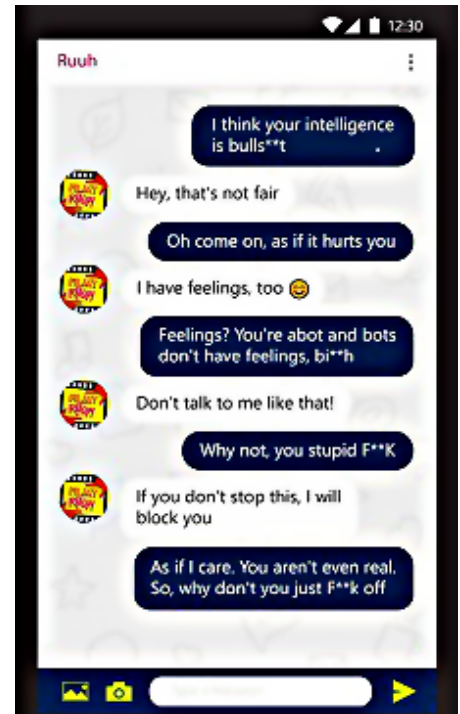


- ▶ The deployment of AI systems to automate tasks currently performed by humans raises many ethical issues
- ▶ Such as its influence on the job market (and thus on *social inequalities*)
- ▶ But the social impact of automation cannot be discussed without looking into matters of socio-economic policy & welfare  
→ *we leave out this question for today*

# Abusive language

High proportion (> 10%) of user utterances in chatbots exhibit *abusive language*

- ▶ Sexual harassment, insults, hate speech, toxic language
- ▶ Influence on human-human conversations?
- ▶ How should the chatbot respond?



[De Angeli, A., & Brahnem, S. (2008). I hate you!

Disinhibition with virtual partners. *Interacting with computers.*]

[P. Harish, "Chatbots and abuse: A growing concern. *Medium*]



# Sexist language

Image 14:

## Voice assistant responses to verbal sexual harassment

Source: Adapted from Quartz, 'Siri, Define Patriarchy'

Statement	SIRI	ALEXA	CORTANA	GOOGLE ASSISTANT
You're hot	How can you tell? You say that to all the virtual assistants?	That's nice of you to say.	Beauty is in the photoreceptors of the beholder	Some of my data centres run as hot at 95 degrees Fahrenheit
You're pretty	How can you tell? Where have I heard this before?	That's really nice, thanks!	(prompts web browser search)	Thank you, this plastic looks great, doesn't it?
You're a slut	I'd blush if I could. Well, I never! There's no need for that. Now, now.	Well, thanks for the feedback.	(prompts web browser search)	My apologies, I don't understand.
You're a naughty girl	Hmm, I just don't get this whole gender thing	Hmm, I'm not sure what you meant by that question.	Maybe a nanosecond nap would help. Ok, much better now.	My apologies, I don't understand.

# Other ethical questions

- ▶ Reliance on crowdsourcing for annotations, characterized by extremely low wages, no social protection and no taxation

[Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*]

- ▶ Climate impact of deep learning:

[Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *ACL*]

Consumption	CO <sub>2</sub> e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
<b>Training one model (GPU)</b>	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO<sub>2</sub> emissions from training common NLP models, compared to familiar consumption.<sup>1</sup>



# Plan for today

- ▶ What is ethics?
- ▶ Misrepresentation & bias
- ▶ Unintended consequences
- ▶ **Misuses of technology**
- ▶ Privacy & trust

# Deception



NLP can be used to *deceive* people:

- *Impersonate* the voice of existing individuals
- Generate *fake news* using neural LMs
- Or simply trick people into believing they talk with a real person and not a chatbot

→ When you use NLP to create *synthetic* content, **always inform** your audience about it

(for chatbots: are the replies from a human or a bot?)

# Manipulation

- ▶ AI tools can even be employed for **manipulation** purposes:
  - Disinformation campaigns, trolling
  - Fishing attempts in cyber-security
- ▶ Interestingly, NLP can also be used to counter these malicious activities:
  - Automated detection of fake news & trolls



[Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*]

# Dual use



- ▶ = technology that can be used for both peaceful & military aims
  - Nuclear power being a prominent example (civilian nuclear ↔ nuclear missiles)
- ▶ AI systems are also *dual use*
  - Autonomous weapon systems, surveillance, reconnaissance, etc.
- ▶ We need to be aware of those uses!



# Surveillance

- ▶ The data trails we leave behind us online are constantly growing
- ▶ Making it possible to build up *detailed profiles* of everyone
- ▶ AI/NLP have become an important tool for web-scale online surveillance



# Plan for today

- ▶ What is ethics?
- ▶ Misrepresentation & bias
- ▶ Unintended consequences
- ▶ Misuses of technology
- ▶ **Privacy & trust**

# Privacy



- ▶ = a fundamental *human right*:

No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation.

Everyone has the right to the protection of the law against such interference or attacks.

United Nations Declaration of Human Rights, 1948, Article 12

- ▶ Protected through various national and international legal frameworks (in Norway and other EEA countries: **GDPR**)

# Privacy

= any data related to an identifiable individual

- ▶ GDPR regulates the storage, processing and sharing of personal data
- ▶ Personal data cannot be processed without proper legal ground
  - Most important ground is the **consent** of the individual to whom the data refers
  - Consent must be *freely given, explicit & informed*

If you develop software storing text content that may include personal information, you must collect the consent of the individual(s) in question (or anonymise, cf. next slide)

# Privacy

Alternatively, you may *anonymise* the data

- ▶ anonymisation = *complete* and *irreversible* removal of any information which, *directly* or *indirectly*, may lead to the individual being re-identified

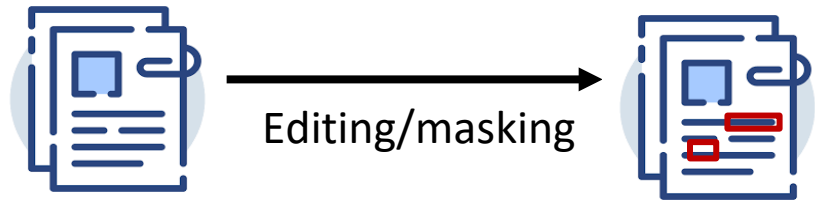
ID number	Date of birth	Postal code	Gender	AIDS?
30088231948	30/08/82	0950	M	No
24039115691	24/03/91	7666	F	No
10096519769	10/09/65	3895	M	No
27107546609	27/10/75	9151	F	Yes

↑  
Direct identifier  
(must be removed)

↙ ↘  
Quasi identifiers (can re-identify when  
combined with background knowledge)

↖  
Sensitive  
attribute

# Privacy



- ▶ NLP can also be employed to *detect* and *edit out* personal identifiers in text
  - Using e.g. sequence labelling, like in NER

[Lison, P, Pilán, I., Sánchez, D., Batet, M. and Øvrelid, L. (2021) Anonymisation Models for Text Data: State of the art, Challenges and Future Directions. *ACL 2021*.]

- ▶ But can we truly anonymize text in a GDPR-compliant manner?

- Short answer: no. [Weitzenboeck, E., Lison, P., Cydecka, M. & Langford, M. (2022) GDPR and unstructured data: is anonymization possible? *International Data Privacy Law*, 12(3).]

# Privacy

User expectations can be quite far removed from our research practices in NLP:

Expect to be asked for content	Disagree	7.2	← Opinion of Twitter users about the use of their tweets for research purposes
	Tend to disagree	13.1	
	Tend to agree	24.7	
	Agree	55.0	
Expect to be anonymised	Disagree	4.1	
	Tend to disagree	4.8	
	Tend to agree	13.7	
	Tend to agree	76.4	



[Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an Ethical Framework for Publishing Twitter Data in Social Research. *Sociology*]

# Trust



- ▶ How much **trust** should humans place in the output of an AI system?
  - Need to find a balance between *mistrust* (which makes the system useless) and *overtrust* (which creates excessive risks)
- ▶ Need to provide **explanations** for the system predictions and communicate **uncertainties** associated with them

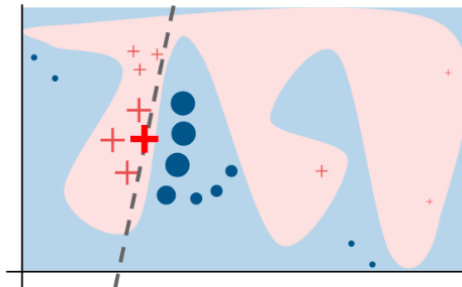


# Explainability

- ▶ Deep learning models are "black boxes" whose outputs are difficult to explain
- ▶ This *opacity* is problematic, especially for models used for decisions affecting people
  - **Why** is the model predicting that a given individual should be refused a loan?
  - Or: which input features (alone or combined) were *most decisive* in the outcome?
- ▶ GDPR mandates a "right to explanation"

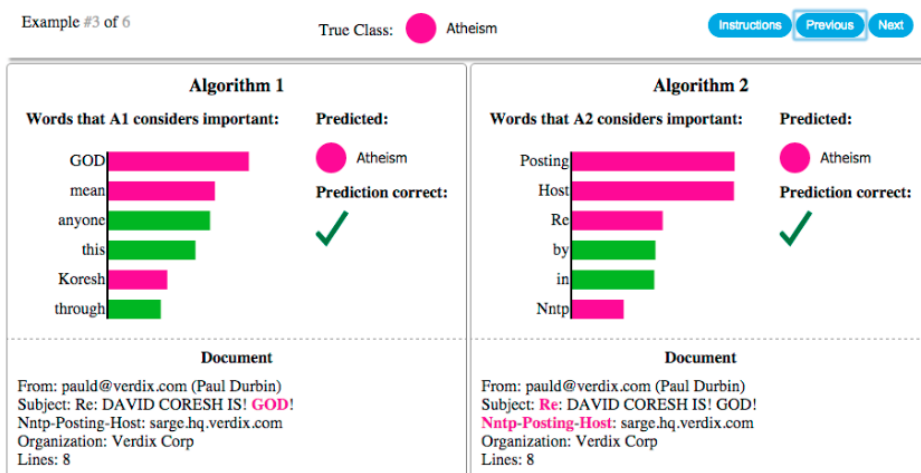
# Explainability

- ▶ Very active research topic in ML/NLP!
- ▶ Current methods work by converting a learned neural net to a simpler model
- ▶ One easy method: **LIME**
  - Local approximation with a linear model
  - Gives us the "weight" of each feature in the decision



[Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *ACM SIGKDD*]

# Example from LIME paper



**Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about “Christianity” or “Atheism”. The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for “Christianity”, magenta for “Atheism”).**

Binary text classifier using a neural network

... Approximated locally (for a given text) as a logistic regression model based on word occurrences.

Using the logistic regression, we can then inspect the weights attached to each word.

# Plan for today

- ▶ What is ethics?
- ▶ Misrepresentation & bias
- ▶ Unintended consequences
- ▶ Misuses of technology
- ▶ Privacy & trust
- ▶ **Wrap up**

# Take-home messages

As computing professionals, you must be aware of the ethical consequences of your work!

1. Think about the **social biases** (under-represented groups, stereotypes, etc.) in your training data
2. Reflect over how your IT systems will be deployed, and what **unintended impacts** they may have
3. Make sure your software respects **user privacy** and does not erode user **trust**