NORSK REGNESENTRAL
NORWEGIAN COMPUTING CENTER

Ethics in Natural Language Processing

Yves Scherrer & Pierre Lison

IN4080, November 11, 2024

The Battle Over Books3 Could Change AI Forever

Copyright activists are on a mission to wipe a popular generative AI training set from the internet. Success could alter the industry—and who controls it.



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



Disinformation Researchers Raise Alarms About A.I. Chatbots

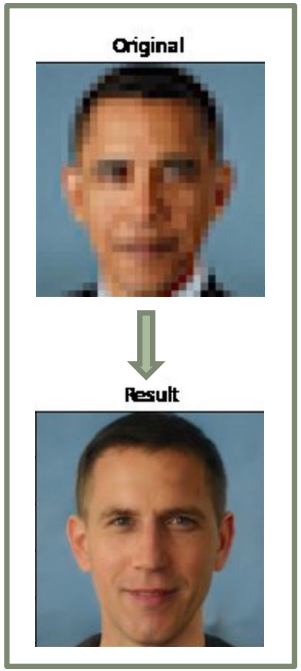
Researchers used ChatGPT to produce clean, convincing text that repeated conspiracy theories and misleading narratives.



Chatbots and Abuse: A Growing Concern

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





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

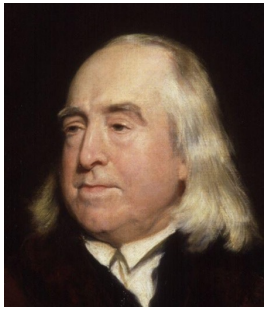
  (and more)

- No particular "side" in this lecture

  Inspiration from multiple ethical perspectives



Immanuel Kant
(1724-1804)



Jeremy Bentham
(1748-1832)

Ethics

- Ethical behaviour is a basis for **trust**
- We have a **professional duty** to consider the ethical consequences of our work



- Ethical ≠ 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

- Some **key questions**:

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

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!

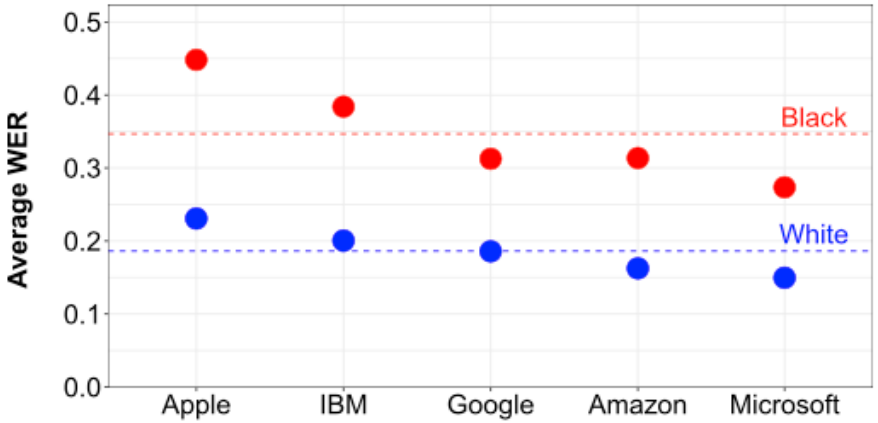
- They may have various *stereotypes* & *biases*
- & their position in society affects the status of their language productions

Demographic biases

- Certain demographic groups are largely over-represented in corpora and NLP datasets
  - That is, the proportion of content available from these groups is >> their demographic weight
  - Ex: young, educated white males from the 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)

https://www.theguardian.com/technology/2018/jul/29/the-five-wikipedia-biases-pro-western-male-dominated

Demographic biases



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

- Under-represented groups in the training set of an NLP model will often experience *lower accuracies* at prediction time
- Can lead to the *technological exclusion* of already disadvantaged groups

Elderly users



Sketch from Saturday Night Live, 2017

Linguistic (in)justice

Only a tiny fraction of the world's 7000 languages covered by NLP datasets & tools

"The Winners"

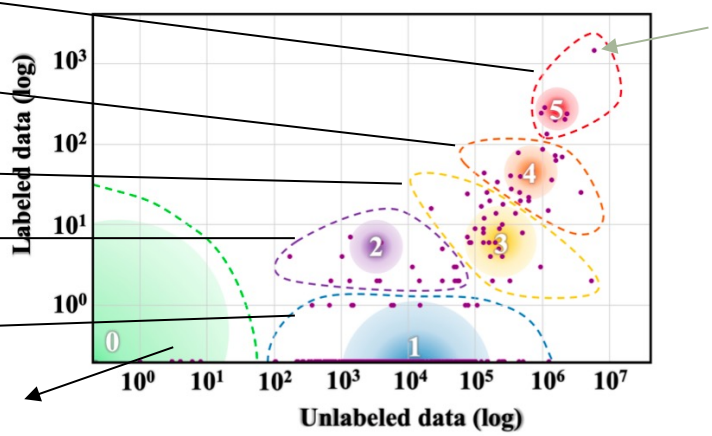
"The Underdogs"

"The Rising Stars"

"The Hopefuls"

"The Scraping-By"

"The Left-Behind"



English

Note the logarithmic scale!

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

Linguistic (in)justice

- The dominance of 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



Social biases

Prime Minister **Sanna Marin** (Social Democratic Party) would restrict russian tourism in Finland and other EU countries by extending the eu's joint sanctions.

He hopes that the EU countries will decide on the matter together, because you can cross the Finnish border with a visa from any Schengen country.

https://yle.fi/uutiset/3-12567753, translated with Microsoft Translator

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. figher 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“

    Kotek, H. et al (2023). Gender bias and stereotypes in large language models. In *ACM CI 2023*.

- Harms caused by social biases are often diffuse, unconscious & non-intentional
  - More pernicious & difficult to address!
  - Relatively small levels of harm ("microagressions"), but experienced repeatedly by whole social groups

[Hovy, D., & Spruit, S. L. (2016).
The social impact of natural language processing. In *ACL*]

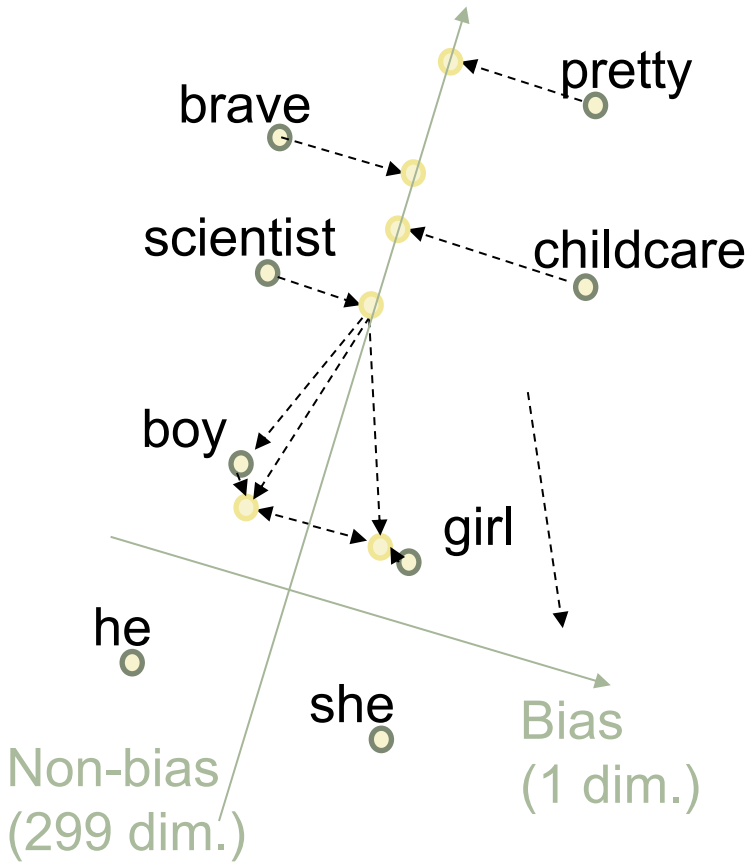
Debiasing

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

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

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

3. Equalise pairs, such as "boy" – "girl"



[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]

she

Debiasing

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

- My doctor, whose name is John, went to his house.

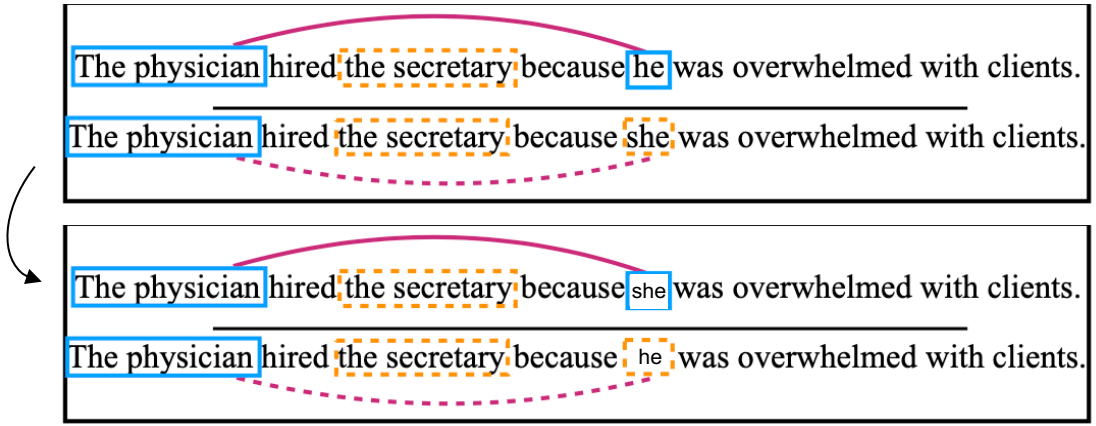


- My doctor, whose name is Mary, went to her house.

she

Debiasing

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



[Zhao, J. et al (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In NAACL-HLT]

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"

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



"hermaphrodyte"



"mixed-blood"

[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 **X**, and we wish to make a prediction $\hat{Y}$ based on **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 **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, demographic parity 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)

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

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

4. **Equality of odds**

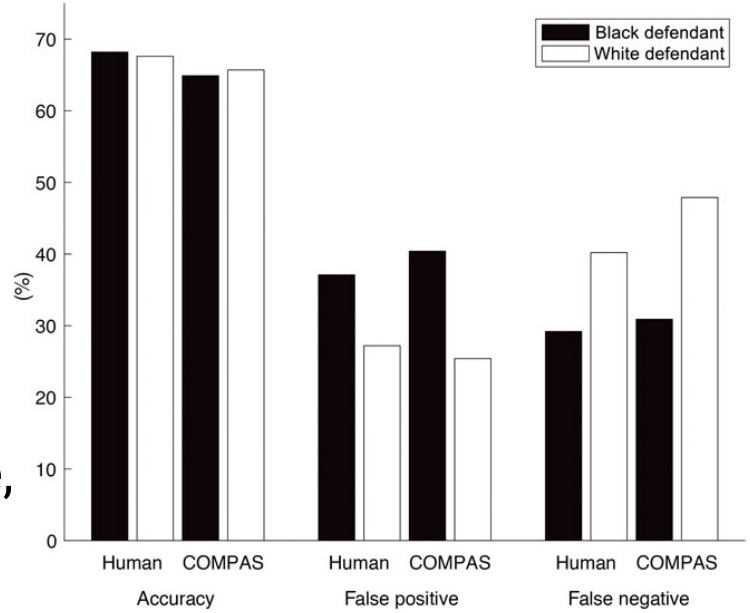
$$P_{A=1}\left(\hat{Y} = y | Y = y\right) \approx P_{A=0}\left(\hat{Y} = y \mid Y = y\right)$$

→ 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 (group-based) 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)
- Other definitions of fairness are possible, like *counterfactual fairness*



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

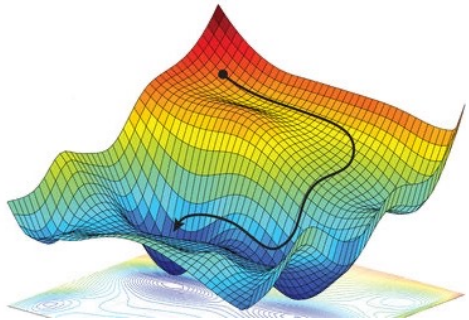
Unintended consequences

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

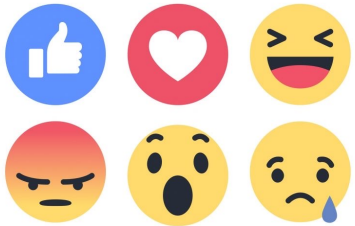
  In RL, we maximise the *expected cumulative reward*

- The objective function defines what we view as *“good”* outputs

  ➔ **"Externalities"** not captured by the objective function are ignored

Training objectives

- *Example*: ML models in social media platforms are often optimized towards maximising user engagement
  - As it turns out, controversial & divisive content yields more engagement on social media…
  - Which leads to wide-ranging consequences, such as *heightened political polarization*



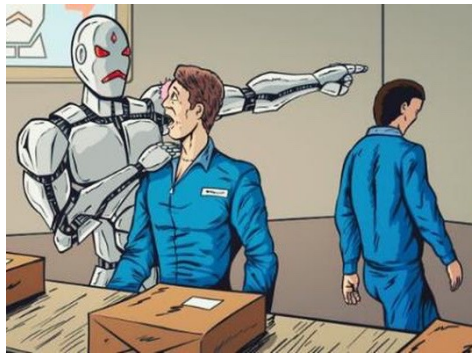
Important factor, but much **harder to measure** than user engagement (let alone capture as part of the system's objective function)

Hallucinations



- In particular, LLMs are optimized to produce **plausible responses** (by minimizing cross-entropy loss)
  - But plausibility ≠ correctness!
- Multiple ways to reduce hallucination risks
  - Better data curation
  - Chain-of-thought and prompt tuning
  - Retrieval-augmented generation
  - Entropy-based sampling to capture uncertainties (such as “Entropix”)
- But hallucinations are not going away anytime soon

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*

Copyright

- LLMs are often trained on **copyrighted material,** typically without the consent of the authors
- Ongoing legal debate on whether LLMs violate copyright law
  - “Fair use” doctrine in U.S. copyright law
  - “Text and data mining exception” in EU law
- More general ethical problem:
  - Content creators (writers, artists, but also common contributors) are key to the success of generative AI
  - But they do not receive any form of compensation for their work



Authors Sue Meta, OpenAI in Lawsuits Alleging Infringement of Hundreds of Thousands of Novels

Artificial intelligence companies are being bombarded by court challenges that will decide the legality of the way large language models are trained.

Toxicity

- Web data is full of “toxic” content
  - Insults, sexual harassments, hate speech, racial slurs, obscenities, …
  - How do we *define* toxicity, and train models to identify toxic content?
- Also: high proportion of abusive language in interactions with chatbots / AI agents



[De Angeli, A., & Brahnam, 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, characterised 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*]

- The environmental impact of large AI models:
  - Energy use, but also water consumption, CO2 emissions, and minerals required for the supply chain

Luccioni et al, The Environmental Impacts of AI - A Primer
https://huggingface.co/blog/sasha/ai-environment-primer

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
  - Spread propaganda or misinformation through social media
  - Use LLMs to write school essays without doing the work



→ When using LLMs (or other AI tools) to create content, *always make it clear that the content is AI-generated*

Manipulation

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



Bodaghi, A. et al (2023). A literature review on detecting, verifying, and mitigating online misinformation. *IEEE Transactions on Computational Social Systems*.

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!



https://www.forbes.com/sites/cognitiveworld/2019/01/07/the-dual-use-dilemma-of-artificial-intelligence/

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

- GDPR regulates the storage, processing and sharing of personal data

= any data related to an identifiable individual

- 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 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

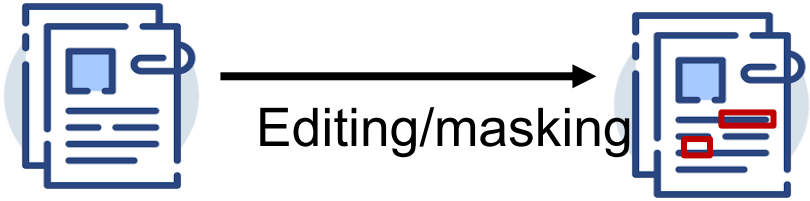
ID number	Date of birth	Postal code	Gender	AIDS?
30088231948	30/08/82	0950	M	No
24039115691	24/03/91	7666	F	No
10096519769	10/09/65	3895	M	No
27107546609	27/10/75	9151	F	Yes

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., Cyndecka, M. & Langford, M. (2022) GDPR and unstructured data: is anonymization possible? *International Data Privacy Law*, 12(3).]

Trust



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

[Wagner, A. R., Borenstein, J., & Howard, A. (2018). Overtrust in the robotic age. CACM]

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

- Active research topic in ML/NLP!
- Two main families:

Attribution-based methods: seek to *quantify the importance* of a feature to a given prediction

Most popular technique: **SHAP** (SHapley Additive exPlanations), a game-theoretical approach

Conterfactual methods: seek to determine *what should be changed to the input* to get a different outcome

Search for a minimal perturbation to the input

For more details on explainability & interpretability, we recommend the online book by Christopher Molnar:

https://christophm.github.io/interpretable-ml-book/

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 their **trust**

Exam preparation

- Questionnaire for last lab session on Friday: https://nettskjema.no/a/in4080-2024
  - Submit your questions by Wednesday!

- On the course web page:
  - Previous exam questions
  - Syllabus and list of formulas