

Ethics

COMP90042

Natural Language Processing

Lecture 22

Semester 1 2021 Week 11
Jey Han Lau



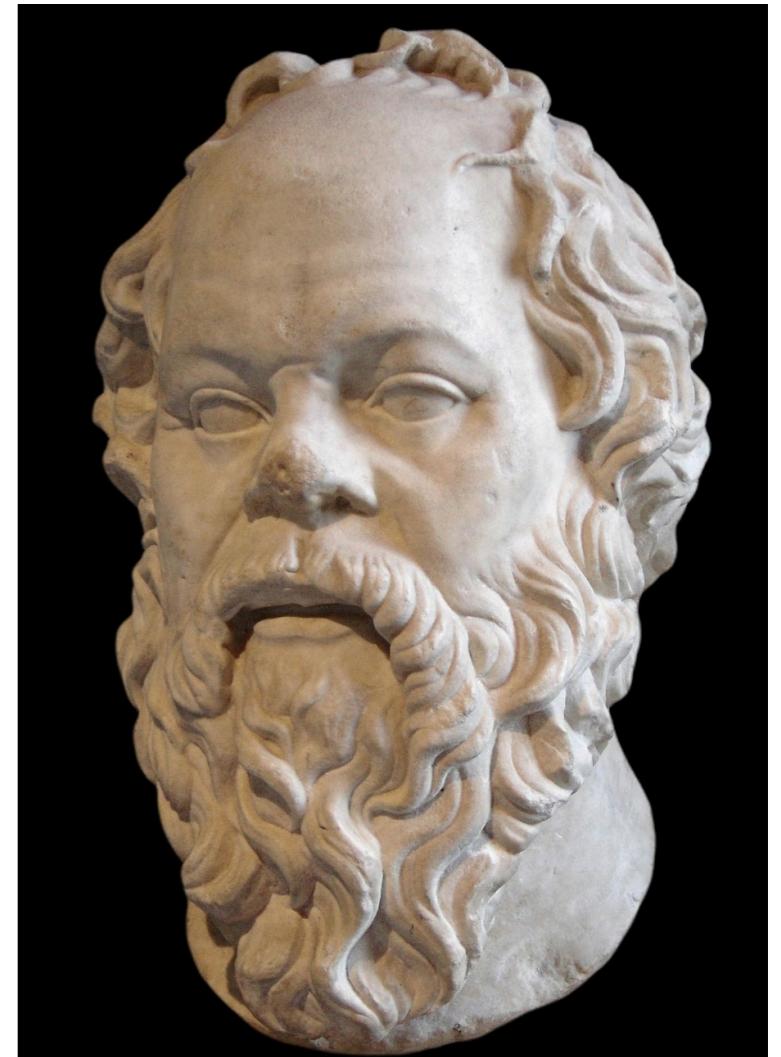
THE UNIVERSITY OF

MELBOURNE

What is Ethics?

How we ought to live – Socrates

- What is the right thing to do?
- Why?



Why Should We Care?

- AI technology is increasingly being deployed to real-world applications
- Have real and tangible impact to people
- Whose responsibility is it when things go bad?



Why Is Ethics Hard?

- Often no objective truth, unlike sciences
- A new philosophy student may ask whether fundamental ethical theories such as utilitarianism is right
- But unlikely a new physics student would question the laws of thermodynamics
- In examining a problem, we need to think from different perspectives to justify our reasons

Learning Outcomes

- Think more about the application you build
 - ▶ Not just its performance
 - ▶ Its social context
 - ▶ Its impact to other people
 - ▶ Unintended harms
- Be a socially-responsible scientist or engineer

Outline

- Arguments against ethical checks in NLP
- Core NLP ethics concepts
- Group discussion

Arguments Against Ethical Checks in NLP

Should We Censor Science?

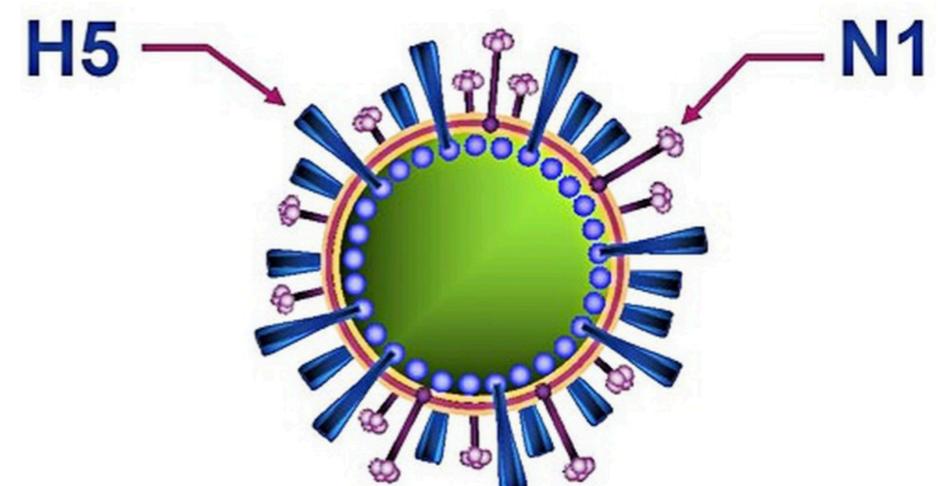
- A common argument when ethical checks or processes are introduced:
 - ▶ Should there be limits to scientific research? Is it right to censor research?
- Ethical procedures are common in other fields: medicine, biology, psychology, anthropology, etc

Should We Censor Science?

- In the past, this isn't common in computer science
- But this doesn't mean this shouldn't change
- Technology are increasingly being integrated into society; the research we do nowadays are likely to be more deployed than 20 years ago

H5N1

- Ron Fouchier, a Dutch virologist, discovered how to make bird flu potentially more harmful in 2011
- Dutch government objected to publishing the research
- Raised a lot of discussions and concerns
- National policies enacted

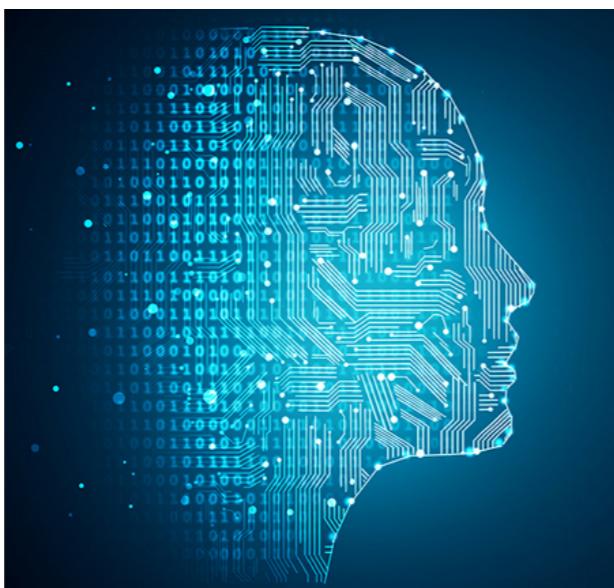


Isn't Transparency Always Better?

- Is it always better to publish sensitive research publicly?
- Argument: worse if they are done underground
- If goal is to raise awareness, scientific publication isn't the only way
 - ▶ Could work with media to raise awareness
 - ▶ Doesn't require exposing the technique

AI vs. Cybersecurity

- Exposing vulnerability publicly is desirable in cyber-security applications
 - ▶ Easy for developer to fix the problem
- But the same logic doesn't always apply for AI
 - ▶ Not easy to fix, once the technology is out



Core NLP Ethics Concepts

Bias

- Most ethics research in NLP focus in this aspect
- A biased model is one that performs unfavourably against certain groups of users
 - ▶ typically based on demographic features such as gender or ethnicity

Bias

- Bias isn't necessarily bad
 - ▶ Guide the model to make informed decisions in the absence of more information
 - ▶ Truly unbiased system = system that makes random decisions
 - ▶ Bad when overwhelms evidence
- Bias can arise from data, annotations, representations, models, or research design

Bias in Word Embeddings

- Word Analogy (lecture 10):
 - ▶ $v(\text{man}) - v(\text{woman}) = v(\text{king}) - v(\text{queen})$
- But!
 - ▶ $v(\text{man}) - v(\text{woman}) = v(\text{programmer}) - v(\text{homemaker})$
 - ▶ $v(\text{father}) - v(\text{mother}) = v(\text{doctor}) - v(\text{nurse})$
 - ▶ Word embeddings reflect and **amplify** gender stereotypes in society
 - ▶ Lots of work done to reduce bias in word embeddings

Dual Use

- Every technology has a primary use, and unintended secondary consequences
 - ▶ nuclear power, knives, electricity
 - ▶ could be abused for things they are not originally designed to do.
- Since we do not know how people will use it, we need to be aware of this duality

OpenAI GPT-2

- OpenAI developed GPT-2, a large language model trained on massive web data (lecture 1 demo)
- Kickstarted the pretrained model paradigm in NLP
 - ▶ Fine-tune pretrained models on downstream tasks (BERT lecture 11)
- GPT-2 also has amazing generation capability
 - ▶ Can be easily fine-tuned to generate fake news, create propaganda

OpenAI GPT-2

- Pretrained GPT-2 models released in stages over 9 months, starting with smaller models
- Collaborated with various organisations to study social implications of very large language models over this time
- OpenAI's effort is commendable, but this is voluntary
- Further raises questions about self-regulation

Privacy

- Often conflated with anonymity
- Privacy means nobody know I am doing something
- Anonymity means everyone know what I am doing, but not that it is me

GDPR

- Regulation on data privacy in EU
- Also addresses transfer of personal data
- Aim to give individuals control over their personal data
- Organisations that process EU citizen's personal data are subjected to it
- Organisations need to anonymise data so that people cannot be identified
- But we have technology to de-identify author attributes

AOL Search Data Leak

- In 2006, AOL released anonymised search logs of users
- Log contained sufficient information to de-identify individuals
 - ▶ Through cross-referencing with phonebook listing an individual was identified
- Lawsuit filed against AOL

Group discussion

Prompts

- Primary use: does it promote harm or social good?
- Bias?
- Dual use concerns?
- Privacy concerns? What sorts of data does it use?
- Other questions to consider:
 - ▶ Can it be weaponised against populations (e.g. facial recognition, location tracking)?
 - ▶ Does it fit people into simple categories (e.g. gender and sexual orientation)?
 - ▶ Does it create alternate sets of reality (e.g. fake news)?

Automatic Prison Term Prediction

- A model that predicts the prison sentence of an individual based on court documents



Automatic CV Processing

- A model that processes CV/resumes for a job to automatically filter candidates for interview

Gandalf the White

Titles

Olórin, Mithrandir, the White Rider, Tharkún, the Grey Pilgrim, Gandalf Greyhame, Stormcrow, Servant of the Secret Fire, Wielder of the Flame of Anor, Gandalf the Grey (ex)

Introductory Statement

An experienced wizard who is looking for a change of career. After travelling extensively over the past 2000 years, I'm looking for a more solid and stable role during my never-ending stay in the undying lands.

I've helped Middle-earth's denizens to combat the powers of evil; now I'm looking for a company to help me combat the power of boredom in the ever-quiet of Valinor. To any role I'll bring wisdom, a sharp tongue and a lot of weed. During power cuts, I'll light the way with my handy staff, errands can be run at super speed on my fast-as-the-wind steed Shadowfax, and of course I'll never be late, arriving precisely when I mean to.

Work History

TA 3018-19	Helped the Free Peoples of Middle-earth defeat the Dark Lord Sauron
TA 3019	Changed robes to White
TA 3019	Killed the Balrog of Moria (Durin's Bane)

Skills & Talents

	Swordmanship
	Use Of Fire And Various Other Spells (Telekinesis And Exorcism)

Language Community Classification

- A text classification tool that distinguishes LGBTQ from heterosexual language
- Motivation: to understand how language used in the LGBTQ community differs from heterosexual community



Take Away

- Think about the applications you build
- Be open-minded: ask questions, discuss with others
- NLP tasks aren't always just technical problems
- Remember that the application we build could change someone else's life
- We should strive to be a socially responsible engineer/scientist

Readings (Optional)

- *The Elements of Moral Philosophy* by James and Stuart Rachels