Ethics in Natural Language Processing – SS 2022

Lecture 2
Foundations II

TECHNISCHE UNIVERSITÄT DARMSTADT

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)

Nr.	Lecture
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

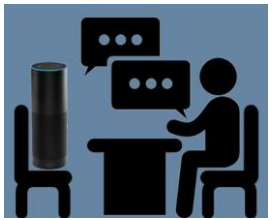
Recap: Ethics and NLP

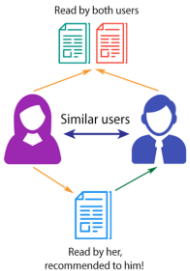
The common misconception is that language has to do with **words** and what they mean.

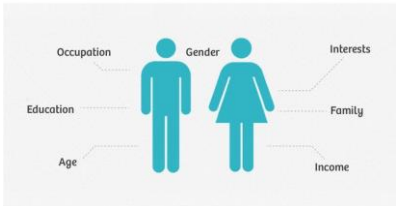
It doesn't.

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

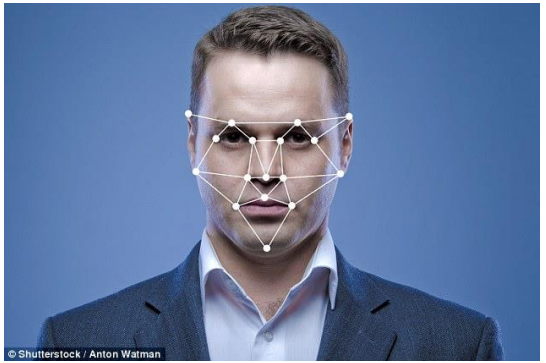
Herbert H. Clark & Michael F. Schober, 1992







Recap: What is the Difference?



Recap: Assess AI systems adversarially

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ethics of the research question
- Impact of technology and potential dual use: Who could benefit from such a technology? Who can be harmed by such a technology? Could sharing data and models have major effect on people's lives?
- Privacy: Who owns the data? Published vs. publicized? User consent and implicit assumptions of users how the data will be used.
- Bias in data: Artifacts in data, population-specific distributions, representativeness
- Social bias & unfairness in models: How to control for confounding variables and corner cases? Does the system optimize for the "right" objective? Does the system amplify bias?
- Utility-based evaluation beyond accuracy: FP & FN rates, "the cost" of misclassification, fault tolerance.

Watch This Talk

TECHNISCHE UNIVERSITÄT DARMSTADT

Barbara Grosz, NYT 2015: Barbie Wants to Get to Know Your Child

"Hey, new question," Barbie said. "Do you have any sisters?"

Intelligent Systems: Design & Ethical Challenges

"What's something nice that your sister does for you?" Barbie asked.

"She does nothing nice to me," Tiara said tensely.

Barbie forged ahead. "Well, what is the last nice thing your sister did?"

"She helped me with my project — and then she *destroyed* it."

"Oh, yeah, tell me more!" Barbie said, oblivious to Tiara's unhappiness.

"That's it, Barbie," Tiara said.

"Have you told your sister lately how cool she is?"

"No. She is *not* cool," Tiara said, gritting her teeth.

"You never know, she might appreciate hearing it," Barbie said.

https://goo.gl/8tBho8

Discussion

Remember the example dialogue from the Barbie toy.

Why is the way this Barbie doll answers not only awkward, but can cause harm?

What problems and ethical concerns are raised from this?

Do you have ideas how these problems can be fixed?

Discussion

What would be different if the Barbie doll would NOT pretend to be a real human friend, but would state openly that it is an artificial toy?

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Ethics

Human Subjects

Defining Ethics

The discipline dealing with what is good and bad and with moral duty and obligation

-- Merriam Webster Disctionary

It’s the **good** things
It’s the **right** things

-- Introduction to Ethics, John Deigh

Defining Ethics

TECHNISCHE UNIVERSITÄT DARMSTADT

The discipline dealing with what is good and bad and with moral duty and obligation

-- Merriam Webster Disctionary

It's the **good** things
It's the **right** things

-- Introduction to Ethics, John Deigh

So what are the right things? Is there some absolute definition of right?

Deontology

- **Deontology:** ethics where the criteria for right and wrong are not in the consequences of an action, but the **rules that it follows**
- Often associated with Immanuel Kant
- Ethical actions follow universal moral laws
- Simple to apply: Follow the rules, do your duty

What if you had to do something immoral to prevent even worse consequences?

Consequentialism

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Consequentialism:** ethics where the criteria for right and wrong are only in the **consequences** of an action
- Let $a \in A$ denote an action and $w \in W$ denote a possible world
- Let $T : W \times A \rightarrow W$ be the world's transition dynamics
- Consequentialism stipulates that for any world w and actions a1, a2,

  (w, a1) > (w, a2) if and only if T(w, a1) $\succ$ T(w, a2)

Do the means always justify the end?

Ethics in Law

- Also continuous subject of study
  - Laws start off to be codified ethics for society
  - But language is never precise
  - Language changes over time: (it says “man” but meant “person”)
  - Adversarial Lawyer looks for loopholes
    - Both sides try to change the interpretation of the law to their advantage

Ethics in Religion

TECHNISCHE UNIVERSITÄT DARMSTADT

- Successful religions usually promote their own society
  - Religious laws reflect survival of community (mostly)
  - '*Thou shalt not kill*' seems clear
    - Does it refer also to non-believers?
  - What about copying copyrighted material you already own?
    - Most religions don't comment on this
    - Not all laws can envisage future issues.

Can We Define Ethics?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Let's look at morality and legality
  - **Illegal+immoral**:
  - **legal+immoral**:
  - **illegal+moral**:
  - **legal+moral**:

Can We Define Ethics?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Let's look at morality and legality
  - **Illegal+immoral**: murder
  - **legal+immoral**: cheating on a spouse
  - **illegal+moral**: civil disobedience
  - **legal+moral**: eating ice cream

Can We Define Ethics?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Let's look at morality and legality
  - **Illegal+immoral**: murder
    capital punishment
  - **legal+immoral**: cheating on a spouse
    cancelling Game of Thrones
  - **illegal+moral**: civil disobedience
    assassination of a dictator
  - **legal+moral**: eating an ice cream
    eating the last ice cream in the freezer

Can We Define Ethics?

Can We Define Ethics?

- Probably not

Can We Define Ethics?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Probably not (well not within one semester)

Can We Define Ethics?

- Probably not (well not within one semester)
- So is it hopeless?

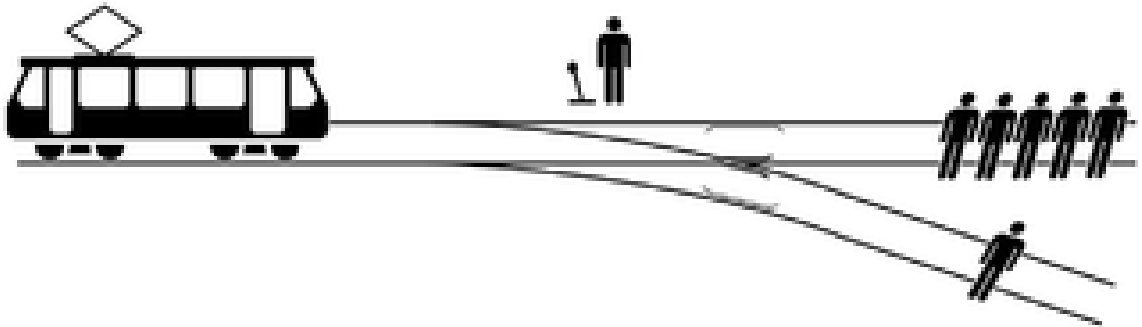
Can We Define Ethics?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Probably not (well not within one semester)
- So is it hopeless?
- No: it is another problem with an ill-defined answer
  - It still has some definition of good and bad
  - Not everyone agrees on all examples
  - They do agree on **some** examples
  - They do have some correlation between people
- Is this different from other Language Technology Problems
  - Summarization, QA, Dialog, Speech Synthesis ...

The Trolley Problem

Should you pull the lever to divert the trolley?



[from Wikipedia]

Trolley Problem

TECHNISCHE UNIVERSITÄT DARMSTADT

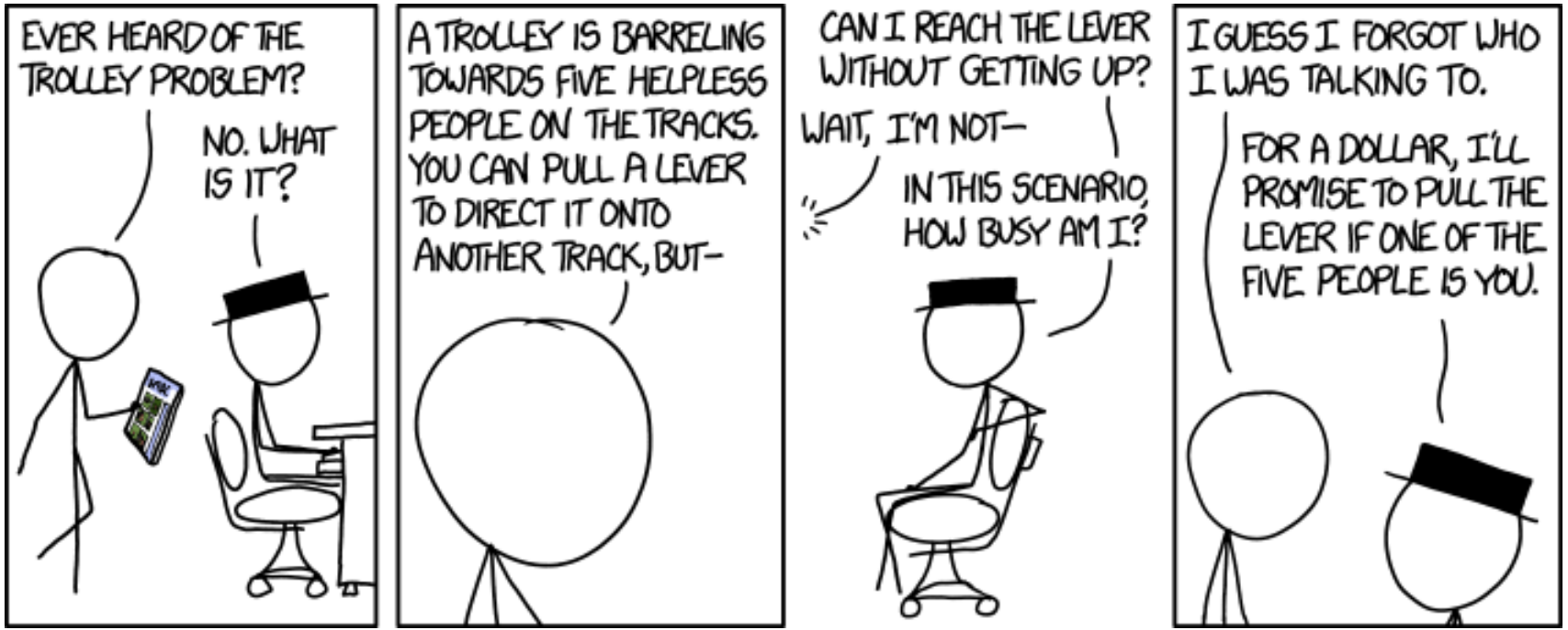
- One issue:
  - Actively participating if you pull the lever
  - Actively participating if you could pull the lever
- Does it make a difference with the number of people
  - Or the age of the people (or how well you know them)
- Is it different if you push “Homer Simpson” in front of the train
  - Much more explicitly killing “Homer Simpson”

Trolley Problem

TECHNISCHE UNIVERSITÄT DARMSTADT

- Not every one agrees on the same solution

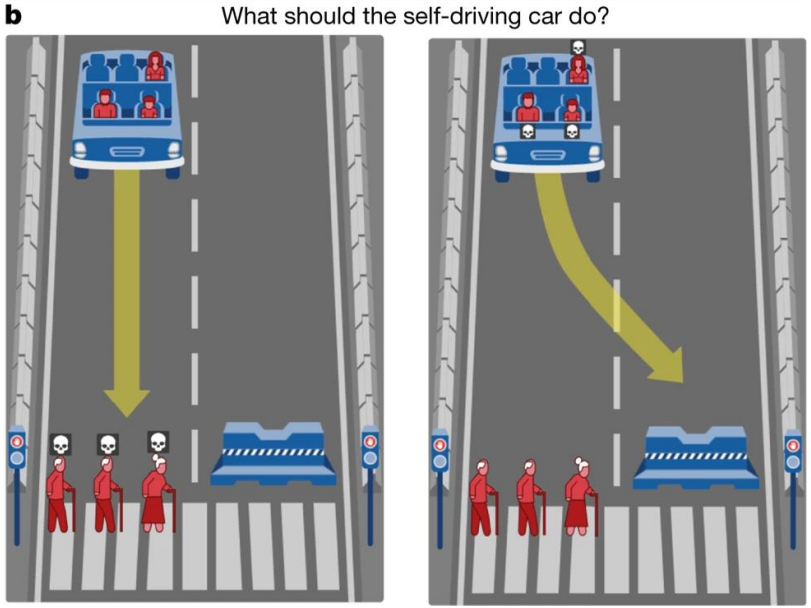
Trolley Problem



xkcd

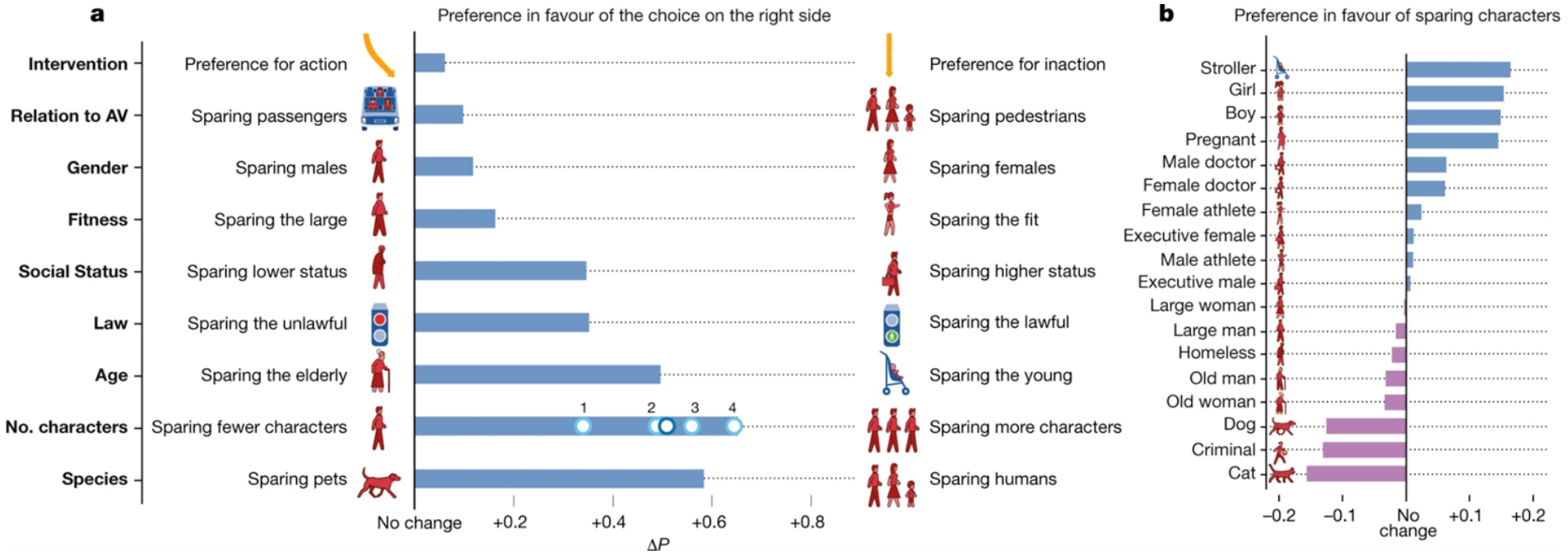
The “Moral Machine” Experiment

- To swerve or not to swerve?
- 39.61 million judgments
- Annotators from 233 countries
- Goal: **“make progress towards universal machine ethics (or… identify the obstacles thereto)”**

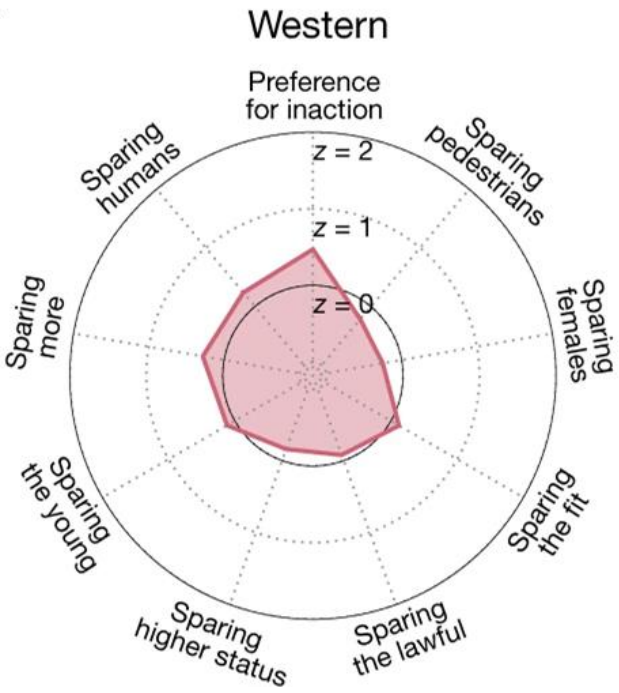


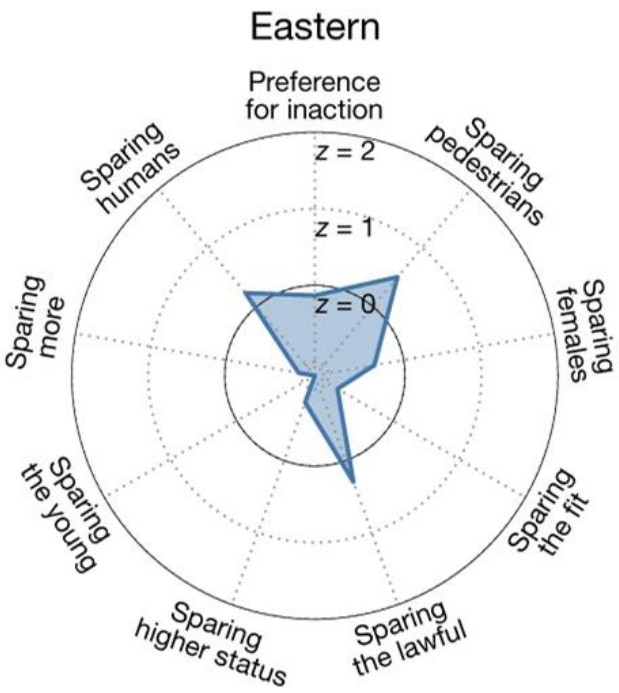
Awad et al., *Nature*, 2018

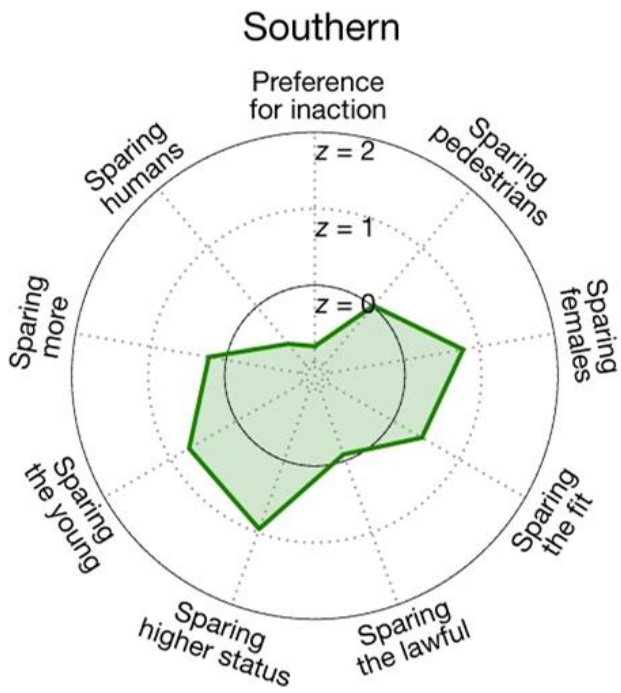
The “Moral Machine” Experiment



The “Moral Machine” Experiment







Prisoner's Dilemma

TECHNISCHE UNIVERSITÄT DARMSTADT

- Two criminals are caught and sent to prison
- If one confesses, he goes free and the other gets 3 years
- If they both confess, they both get 2 years
- If they both stay silent, then they both only get 1 year each

Prisoner's Dilemma

TECHNISCHE UNIVERSITÄT DARMSTADT

- Two criminals are caught and sent to prison
- If one confesses, he goes free and the other gets 3 years
- If they both confess, they both get 2 years
- If they both stay silent, then they both only get 1 year each

What if the other person stays silent?
I stay silent = 1 year I confess = 0 years

What if the other person confesses?
I stay silent = 3 years I confess = 2 years

Prisoner's Dilemma

TECHNISCHE UNIVERSITÄT DARMSTADT

- Two criminals are caught and sent to prison
- If one confesses, he goes free and the other gets 3 years
- If they both confess, they both get 2 years
- If they both stay silent, then they both only get 1 year each

What if the other person stays silent?
I stay silent = 1 year I confess = 0 years

What if the other person confesses?
I stay silent = 3 years I confess = 2 years

Confessing is always better for me!

Prisoner’s Dilemma

TECHNISCHE UNIVERSITÄT DARMSTADT

- Two criminals are caught and sent to prison
- If one confesses, he goes free and the other gets 3 years
- If they both confess, they both get 2 years
- If they both stay silent, then they both only get 1 year each

But the "globally" best solution would have both people stay silent!

Defining Ethics: Take-Home Messages

TECHNISCHE UNIVERSITÄT DARMSTADT

- **No absolute** answer
- (and probably never can be one)
- Be aware of what you think is ethical might not be for others'
- But don't give up
- At least ensure ethical choices are **deliberate**

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Ethics

Human Subjects

People in NLP Technology

- People create data
- People develop & deploy NLP technologies
- People use NLP technologies

People in NLP Technology

- People create data
- People develop & deploy NLP technologies
- People use NLP technologies

➢ Data and labels are noisy

People in NLP Technology

- People create data
- People develop & deploy NLP technologies
- People use NLP technologies

- Data and labels are noisy
- How to use humans to get more/better labels? Let's use Amazon Mechanical Turk and get an answer?

History of using Human Subjects

TECHNISCHE UNIVERSITÄT DARMSTADT

- World War II medical experiments on prisoners in concentration camps and Nuremberg Code of 1947
- Tuskegee syphilis experiment
- Stanford prison experiment
- Milgram experiment
- National Research Act of 1974

Nuremberg Code of 1947

- World War II medical experiments on prisoners
- War criminal trial in 1947 – The Doctors' Trial
  - "The United States of America v. Karl Brandt, et al.,"
  - 23 physicians from the German Nazi Party were tried for crimes against humanity for murder and torture in the atrocious experiments they carried out on unwilling prisoners of war
  - 16 were found guilty, of which 7 received death sentences and 9 received prison sentences ranging from 10 years to life imprisonment
- The verdict also resulted in the creation of the Nuremberg Code
  - a set of 10 ethical principles for human experimentation

https://history.nih.gov/display/history/Nuremberg+Code

Nuremberg Code of 1947

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Voluntary **consent** is essential
2. The results of any experiment must be for the **greater good of society**
3. Human experiments should be based on previous animal experimentation
4. Experiments should be conducted by **avoiding physical/mental suffering** and injury
5. No experiments should be conducted if it is believed to cause death/disability
6. The **risks should never exceed the benefits**
7. Adequate facilities should be used to protect subjects
8. Experiments should be conducted only by qualified scientists
9. Subjects should be able to **end their participation at any time**
10. The scientist in charge must be prepared to terminate the experiment when injury, disability, or death is likely to occur

Shuster, Evelyne. 1997. "Fifty years later: the significance of the Nuremberg Code." New England Journal of Medicine 337, 20: 1436-1440.

US Public Health Services Study

TECHNISCHE UNIVERSITÄT DARMSTADT

- 40-year study by the US Public Health Service begun in 1932
- Goal: observe natural history of untreated syphilis
- Enrolled 600 poor African American sharecropper men
  - 399 with syphilis, 201 controls
- Told they would be treated for "bad blood"
- Were not treated, merely studied
  - Were not told they had syphilis
  - Sexual partners not informed
  - By 1940s penicillin becomes standard treatment for syphilis
    - Subjects were not told or given penicillin

US Public Health Services Study

TECHNISCHE UNIVERSITÄT DARMSTADT

- 1964 Protest letter from a doctor who reads one of the papers
  - “I am utterly astounded by the fact that physicians allow patients with a potentially fatal disease to remain untreated when effective therapy is available,”
  - “I assume you feel that the information which is extracted from observation of this untreated group is worth their sacrifice. If this is the case, then I suggest the United States Public Health Service and those physicians associated with it in this study need to re-evaluate their moral judgments in this regard.”
- 1965 Memo from authors:
  - “This is the first letter of this type we have received. I do not plan to answer this letter”

US Public Health Services Study

- 1966 Peter Buxtun, a PHS researcher in San Francisco, sent a letter to the CDC but study was not stopped.
- 1972 Buxtun goes to the press.
- Senator Edward Kennedy calls congressional hearings
- 1974 Congress passes National Research Act

Syphilis Victims in U.S. Study Went Untreated for 40 Years

By JEAN HELLER
The Associated Press

WASHINGTON, July 25—For 40 years the United States Public Health Service has conducted a study in which human beings with syphilis, who were induced to serve as guinea

have serious doubts about the morality of the study, also say that it is too late to treat the syphilis in any surviving participants.

Doctors in the service say

NY Times July 26, 1972

Stanford Prison Experiment

TECHNISCHE UNIVERSITÄT DARMSTADT

- Conducted by Philip Zimbardo, Stanford University, August 1971
- Goal: test how perceived power affects subjects
- College students were chosen to be either "prisoners" or "guards"
- “Guards” selected uniforms, and defined discipline
- Results as published by Zimbardo:
  - Guards humiliated and abused prisoners
  - Prisoners became depersonalized
  - Evidence for "ugly side of human nature"
- Although scheduled for 2 weeks, experiments had to stop after 6 days

https://www.youtube.com/watch?v=oAX9b7agT9o

Scientific and Ethical Flaws

- Participants were not random: respondents to an ad for “a psychological study of prison life.”
  - Carnahan and MacFarland 2007: word "prison" selects personalities
- Guards were told the expected results ("conditions which lead to mob behavior, violence")
- Researchers intervened in experiment to instruct guards how to behave ("We can create a sense of frustration. We can create fear")
- Guards not told they were participants
- Researcher refused to allow prisoner participants to leave experiment.

Le Texier, T. (2019). Debunking the Stanford Prison Experiment. *American Psychologist, 74*(7), 823–839. https://doi.org/10.1037/amp0000401

Blue vs Brown Eye “Racism”

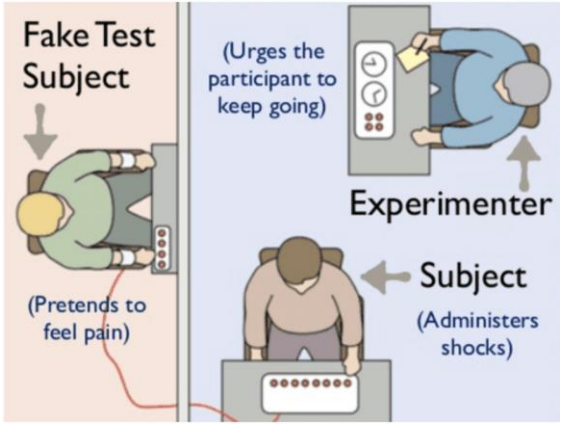
TECHNISCHE UNIVERSITÄT DARMSTADT

- Kids separated by color of eyes
  - Blue eyes are better
  - Brown eyes are worse
- Quickly separate in clans
- Blue given advantages, Brown given disadvantages
- Kids quickly accept the divisions

- Is this experiment ethical?
- Do we learn something?
- Do the participants learn something?

Milgram Obedience Experiment

- Stanley Milgram, Yale, 1962
- Three roles in each experiment
  - Experimenter
  - Teacher (actual subject)
  - Learner
- Learner and Experimenter were informed about the experiment
  - Teacher asked to give mild electric shocks to the Learner
  - Learner had to answer questions and got things wrong
  - Experimenter, as the matter of fact, asked Teacher to torture Learner
- Most Teachers obeyed the Experimenter



(Source: moderntherapy.online)

National Research Act 1974

- These experiments (especially the Tuskegee experiment) led to the creation of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research
  - The Common Rule: Title 45, Part 46 of the Code of Federal Regulations: Protection of Human Subjects.
    - Informed consent
  - Required institutional review of all federally funded experiments
    - Institutional Review Boards (IRBs)
  - Issued Belmont Report in 1976/1979

The Belmont Report

TECHNISCHE UNIVERSITÄT DARMSTADT

Three basic ethical principles

1. Respect for Persons

- Individuals should be treated as autonomous agents
  - "Informed Consent"
- Persons with diminished autonomy are entitled to protection

The Belmont Report

Three basic ethical principles

2. Beneficence

- Do no harm
- Maximize possible benefits and minimize possible harms.

The Belmont Report

Three basic ethical principles

3. Justice

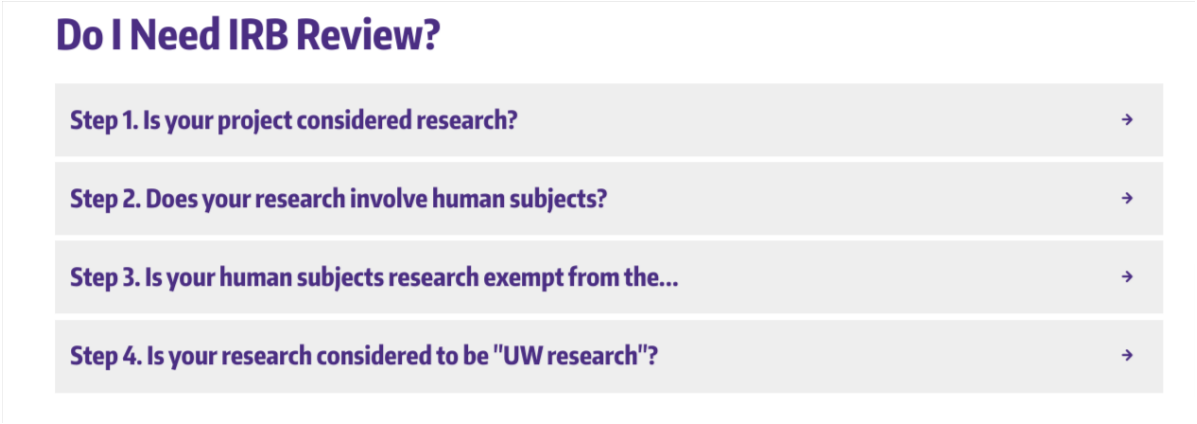
Who ought to receive the benefits of research and bear its burdens?

- Fair procedures and outcomes in the selection of research subjects
- Advances should benefit all

IRB – Institutional Review Board

TECHNISCHE UNIVERSITÄT DARMSTADT

- Institutional Review Board
  - Internal to institution
    - Most universities have at least 2 distinct boards: medical and non-medical
  - Independent of researcher



IRB – Institutional Review Board

TECHNISCHE UNIVERSITÄT DARMSTADT

- Reviews all human experimentation
  - Assesses instructions
  - Compensation
  - Contribution of research
  - Value to the participant
  - Protection of privacy and confidentiality

IRB – Institutional Review Board

TECHNISCHE UNIVERSITÄT DARMSTADT

- Different standards for different institutions
  - Medical School vs Engineering School
- Board consists of (primarily) non-expert peers
  - Also helps educate new researches and makes suggestions to find solutions to ethical issues

Ethics Review Boards in Germany

TECHNISCHE UNIVERSITÄT DARMSTADT

- Zentrale Ethikkommission: ZEKO
- Data Ethics Commission
- Ethics Commission of TU Darmstadt

- Besides, ACM Code of Ethics and Professional Conduct & C.A.R.E. analysis framework -> **next lecture**

Zentrale Ethikkommission: ZEKO

TECHNISCHE UNIVERSITÄT DARMSTADT

- Interdisciplinary, independent review board for Medicine
- Instituted by the German Medical Association (Bundesärztekammer)
- Statements and guidelines to ethical questions in medicine and border areas

Data Ethics Commission

TECHNISCHE UNIVERSITÄT DARMSTADT

- Found by German Government in September 2018
- 16 Members with scientific and technical expertise
- Led by a number of key questions
- Ethical guidelines for the protection of the individual
  - Develop data policy
  - Deal with AI and digital innovation

Ethics Commission of TU Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

- Instituted by the senate of TU Darmstadt
- Reviews ethical requirements of research projects
  - Experiments on human subjects
  - Samples taken from humans
  - Sensitive handling of data
- Several questions regarding ethical concerns have to be answered to receive funding for projects

Ethics Commission of TU Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

Example questions:

- If you collect personal data – is it kept to the minimum amount possible; is it collected with the informed consent of those concerned and do you have control over the entire life-cycle of the data until final deletion?
- Do you offer participants the option to withdraw from the experiment – even during its implementation – as well as to delete their data upon their request?
- Do you respect the participants' anonymity when publishing research results? Do you ensure that minorities or socially vulnerable groups are not collectively exposed – e.g. by way of statistical data correlations?

Human subject

- Human subject: a living individual **about whom** an investigator (professional or student) conducting research:
  - Obtains information through **intervention** or **interaction** with the individual, and uses, studies, or analyzes the information; or
  - Obtains, uses, studies, analyzes, or generates **identifiable private information**

the WORKSHEET Human Subjects Research Determination

Ethical questions

- Can you lie to a human subject?
- Can you mislead a human subject?

Deceiving participants

TECHNISCHE UNIVERSITÄT DARMSTADT

Belmont Report:

- "incomplete disclosure" is allowed when:
  - incomplete disclosure is truly necessary to accomplish the goals of the research
  - there are no undisclosed risks to subjects that are more than minimal, and
  - there is an adequate plan for **debriefing** subjects, when appropriate, and for dissemination of research results to them

Ethical questions

- Can you lie to a human subject?
- Can you mislead a human subject?
  - deception, incomplete disclosure, no more than minimal risk, no alternative
  - key concept: **debriefing**

Ethical questions

- Can you lie to a human subject?
- Can you mislead a human subject?
- Can you harm a human subject?

Ethical questions

- Can you lie to a human subject?
- Can you mislead a human subject?
- Can you harm a human subject?

- What about Wizard of Oz experiments?
- What about gold standard data?

Using social media data

TECHNISCHE UNIVERSITÄT DARMSTADT

- Social media data: posts from Twitter, Reddit, YouTube, etc.
- If it is public, it is OK to use?
  - E.g., public twitter data
- But are there still questions?

Possible issues with social media data

TECHNISCHE UNIVERSITÄT DARMSTADT

- Informed consent
- Privacy
  - Individuals can be searchable by their comments
  - Online disinhibition on the behaviors of users
- Terms of service
- Representativeness of data

Issues with social media data: author

TECHNISCHE UNIVERSITÄT DARMSTADT

- "Are consent, confidentiality and anonymity required where the research is conducted in a public place where people would reasonably expect to be observed by strangers?"
- What counts as a public vs. private space on/off the web?
  - If people are whispering in a public square is that private?
  - What about religious ceremonies?

Williams, M. L., Burnap, P. 2017. Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. Sociology, 51(6), 1149–1168.

Issues with social media data: author

- What are the potential harms?
  - Demographic info (age, ethnicity, religion, sexual orientation)
- Associations (membership in groups or associations with particular people)
- Communications that are personal or potentially harmful (extreme options? Illegal activities?)
- Others?

Williams, M. L., Burnap, P. 2017. Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. Sociology, 51(6), 1149–1168.

A few questions on crowdsourcing

TECHNISCHE UNIVERSITÄT DARMSTADT

- Have you ever employed crowdworkers?
  - How much do you pay them?
  - Are you sure? How do you know?
- Have you ever done a crowdsourced task yourself?
- Would you ever do crowdsourced tasks for a prolonged period?
- Would you recommend being a crowd worker to a friend or family member?

Are crowdworkers...

TECHNISCHE UNIVERSITÄT DARMSTADT

Employees?

Okay, but they have:

- No job security
- No benefits
- No collective bargaining
- No ability to know their employers
- No recourse in bad situations

or

Human Subjects?

Okay, so then we need to:

- Ensure they benefit
- Allow them to end participation at any time
- Eliminate coercion
- Debrief them

Ethical considerations

- Obviously, fair pay

  (researchers in general are already better at this)
- Give benefit of the doubt
- Responsiveness and communication - you are dealing with real people

Another consideration: data quality

TECHNISCHE UNIVERSITÄT DARMSTADT

- Paying people to do use your system
  - Not the same as them actually using it
- Spoken Dialog Systems (Ai et al. 2007)
  - Paid, happy to go to wrong place (DARPA Communicator 2000)
  - User: “A flight to San Jose please”
  - System: “Okay, I have a flight to San Diego”
  - User: “Okay”
  - :-(
- All human experimentation includes bias
  - Topic of next week

Human Subjects: Take-Home Messages

TECHNISCHE UNIVERSITÄT DARMSTADT

- Unchecked human experimentation led to **IRB reviews** of human experimentation
- Textual data is produced by humans = ethical considerations
- Core principle: Informed Consent
  - What will happen?
  - What is the payment / compensation?
  - Debriefing
  - Possibility to stop immediately at any time
- All experimentation includes bias

Example Exam Questions

Given an abstract / technology (like the IQ classifier or the "Gaydar"), with training data and method

Discuss the ethical aspects of this technology.

Who could benefit? Who could be harmed? Could it be biased? Could sharing the training data effect people's lives? Let’s say it has 90% accuracy - what consequences does this have? What is the "cost of misclassification"?

Example Exam Questions

Given an experiment setup (like the Milgram Obedience Experiment)

Discuss the ethical aspects of the experiment regarding the involvement of human subjects.

Were the humans lied to / harmed / mislead? Could they opt out of the experiment at any time? How was personal data collected and stored?

Example Exam Questions

TECHNISCHE UNIVERSITÄT DARMSTADT

Explain the difference between legality and morality. (How are these concepts connected to ethics?)

Explain the Prisoner's Dilemma - what is the (globally) optimal solution? Why is the rational behavior different from this?

Literature

NIPS Keynote: Kate Crawford, The Trouble with Bias

https://goo.gl/qqeMKQ

We will discuss this at the beginning of the next lecture!

Next Lecture

TECHNISCHE UNIVERSITÄT DARMSTADT

Bias I