

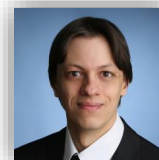
# Ethics in Natural Language Processing – SS 2022



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

## Lecture 1 Introduction + Foundation I

**Dr. Thomas Arnold**  
**Aniket Pramanik**



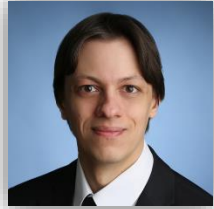
**Ubiquitous Knowledge Processing Lab**  
**Technische Universität Darmstadt**

*Slides and material from Yulia Tsvetkov*



**Carnegie Mellon University**  
Language Technologies Institute

# Introduction: Lecturers



**Dr. Thomas Arnold**



**Aniket Pramanik**

# Outline

**UKP Lab: profile and projects**

**Administrative course issues**

**Introduction to Ethics in NLP**



UKP

## Ubiquitous Knowledge Processing

Lexical-  
semantic  
algorithms

Lexical-  
semantic  
resources

Semantic  
search

Statistical  
semantics

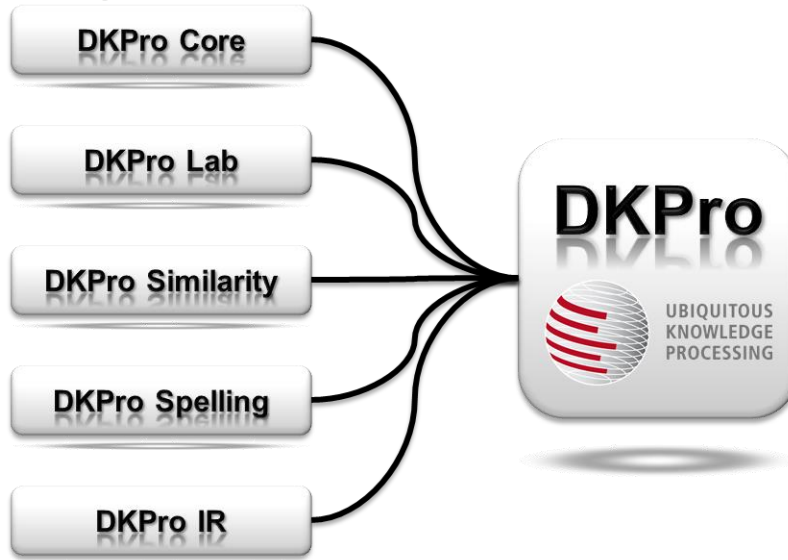
Educational  
language  
technology

Language  
technology  
for  
eHumanities

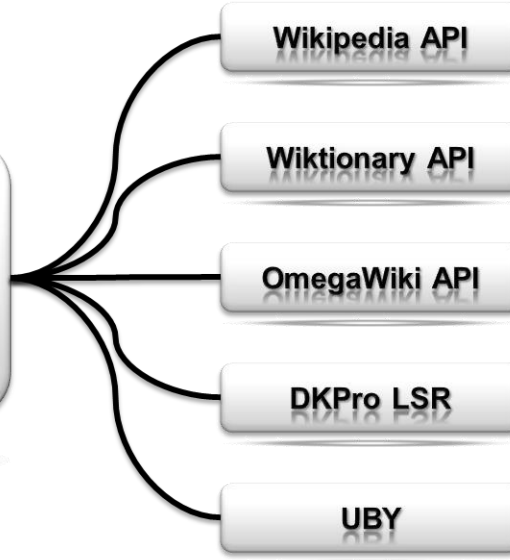
UKP Lab conducts foundational research in computational linguistics and language technology with a focus on lexical-semantic processing.

We study novel applications in the area of educational research, digital humanities, and information management and put a focus on the scarcely researched discourse types in the Web.

## Processing



## Resources



[https://www.informatik.tu-darmstadt.de/ukp/research\\_6/software\\_3/](https://www.informatik.tu-darmstadt.de/ukp/research_6/software_3/)

# Teaching Concept – University

Audience of UKP lectures:

- B.Sc. / M.Sc. / Diploma Informatik
- M.Sc. “Internet- und Web-basierte Systeme”
- M.Sc. Wirtschaftsinformatik
- B.Sc. / M.Sc. „Psychologie in IT“
- Joint Bachelor of Arts
- M.A. Linguistic and Literary Computing
- PhD
- Nebenfach Informatik (multiple FBs)

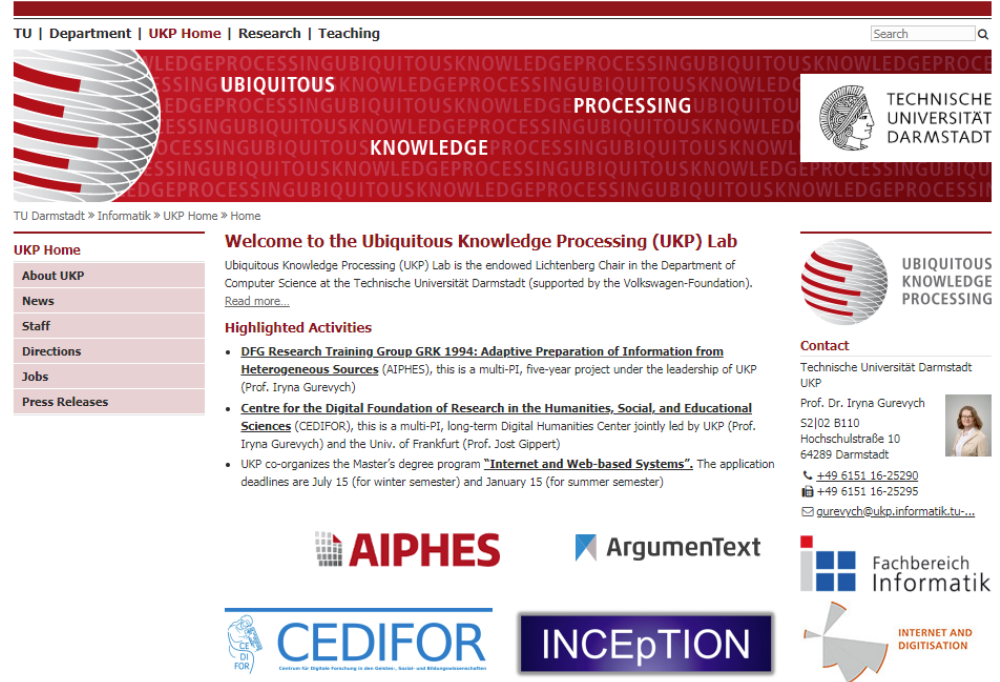
Interdisciplinary nature of this lecture!

# More Information?

See our website:  
[www.ukp.tu-darmstadt.de](http://www.ukp.tu-darmstadt.de)



UBIQUITOUS  
KNOWLEDGE  
PROCESSING



The screenshot shows the homepage of the Ubiquitous Knowledge Processing (UKP) Lab. The header includes navigation links for TU, Department, UKP Home, Research, and Teaching, along with a search bar. The main banner features a repeating pattern of the words 'UBIQUITOUS KNOWLEDGE PROCESSING' in red and white. Below the banner, the 'UKP Home' section contains a sidebar with links to 'About UKP', 'News', 'Staff', 'Directions', 'Jobs', and 'Press Releases'. The main content area welcomes visitors to the UKP Lab, which is the endowed Lichtenberg Chair in the Department of Computer Science at the Technische Universität Darmstadt. It lists 'Highlighted Activities' including the DFG Research Training Group GRK 1994: Adaptive Preparation of Information from Heterogeneous Sources (AIPHES), the Centre for the Digital Foundation of Research in the Humanities, Social, and Educational Sciences (CEDIFOR), and the UKP co-organization of the Master's degree program 'Internet and Web-based Systems'. The 'Contact' section provides the address of the UKP Lab, contact information for Prof. Dr. Iryna Gurevych, and a photo of her. At the bottom, logos for partner organizations are displayed: AIPHES, ArgumenText, CEDIFOR, INCEpTION, and the Fachbereich Informatik.

TU | Department | **UKP Home** | Research | Teaching

Search

TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

TU Darmstadt » Informatik » UKP Home » Home

**UKP Home**

- About UKP
- News
- Staff
- Directions
- Jobs
- Press Releases

**Welcome to the Ubiquitous Knowledge Processing (UKP) Lab**

Ubiquitous Knowledge Processing (UKP) Lab is the endowed Lichtenberg Chair in the Department of Computer Science at the Technische Universität Darmstadt (supported by the Volkswagen-Foundation).  
[Read more...](#)

**Highlighted Activities**

- **DFG Research Training Group GRK 1994: Adaptive Preparation of Information from Heterogeneous Sources** (AIPHES), this is a multi-PI, five-year project under the leadership of UKP (Prof. Iryna Gurevych)
- **Centre for the Digital Foundation of Research in the Humanities, Social, and Educational Sciences** (CEDIFOR), this is a multi-PI, long-term Digital Humanities Center jointly led by UKP (Prof. Iryna Gurevych) and the Univ. of Frankfurt (Prof. Jost Gippert)
- UKP co-organizes the Master's degree program **"Internet and Web-based Systems"**. The application deadlines are July 15 (for winter semester) and January 15 (for summer semester)

**Contact**

Technische Universität Darmstadt  
UKP

Prof. Dr. Iryna Gurevych

52102 B110  
Hochschulstraße 10  
64289 Darmstadt

+49 6151 16-25290  
+49 6151 16-25295  
[gurevych@ukp.informatik.tu-darmstadt.de](mailto:gurevych@ukp.informatik.tu-darmstadt.de)

**AIPHES**

**ArgumenText**

**CEDIFOR**  
Center for Digital Foundation in the Humanities, Social, and Educational Sciences

**INCEpTION**

**Fachbereich Informatik**

**INTERNET AND DIGITISATION**

# Warm up

- Hands up
  - Computer Science?
  - Other disciplines?
  - Bachelor?
  - Master?
- Other UKP Lectures?
  - Foundations of NLP / FOLT
  - Text Analytics Seminar
  - NLP and the Web



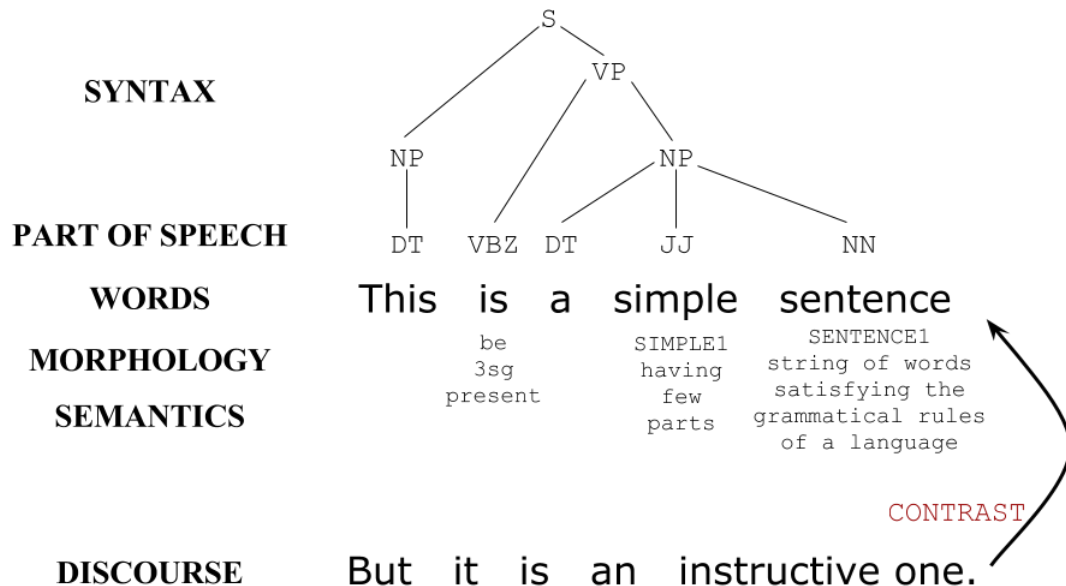
# Outline

**UKP Lab: profile and projects**

**Administrative course issues**

**Introduction to Ethics in NLP**

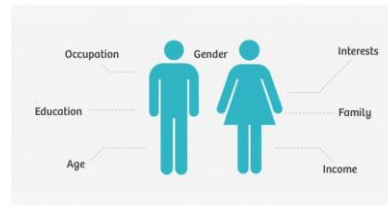
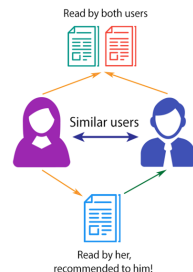
# What NLP Has To Do With Ethics?



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



# Course Goals

- Explain philosophical and practical aspects of ethics
- Show the limits and limitations of machine learning models
- Use techniques to identify and control bias and unfairness in models and data
- Demonstrate and quantify the impact of influencing opinions in data processing and news

# Disclaimer

- We are no Ethics and Philosophy experts!
- Therefore, our approach on philosophical foundations will be quite shallow compared to a Philosophy class
- We will discuss ethical aspects in example cases of NLP and Machine Learning applications

# General Information

- There will be a final exam for the Ethics in NLP lecture
- The practice class contains 5 homework exercises
- The lecture slides, handouts, readings etc. can all be found on the Moodle e-Learning platform:
  - <https://moodle.informatik.tu-darmstadt.de>
  - No password is required
  - course name: **EthicsNLP22**

## No fixed date yet

- **Where:** ???
- **Relevant registration times as usual**
- **Content:** lecture, readings, practice class

We will include 1 – 2 questions from the practice class in the exam  
(Some Python code, questions about a practice class topic...)

- Practice Class starts this week: 5 homework assignments
  - computational **analysis** of newspaper corpora
  - **detection** of bias in directed speech
  - **classification** of different types of bias or hate speech
  - **generation** of paraphrases that anonymize or obfuscate demographic properties of a writer
- Programming language: Python (we have a tutorial for you, if needed)
- With at least 75 % of the points, you can improve your exam grade by 0.3 (or 0.4). You have to pass the exam without the bonus!



# Questions

**Any questions, suggestions...?**

[arnold@ukp.informatik.tu-darmstadt.de](mailto:arnold@ukp.informatik.tu-darmstadt.de) (Lecture)

[pramanik@ukp.informatik.tu-darmstadt.de](mailto:pramanik@ukp.informatik.tu-darmstadt.de) (Practice class)

Or use the moodle forum!

# Topics Covered

- Foundations
- Misrepresentation and bias
- InCivility in communication, hate speech
- Privacy and Security
- Democracy and the language of manipulation
- NLP for Social Good

## Part 1: Foundations

- What is ethics
- History
- Medical and psychological experiments
- IRB and human subjects

## Part 2: Misrepresentation and bias

- Theoretical background, IAT
- Algorithms to identify bias in NLP models and data
- Debiasing

## Part 3: InCivility in communication

- Techniques to monitor trolling, hate speech, abusive language, cyberbullying, toxic comments
- Hate speech and bias in conversational agents

## Part 4: Privacy and Security

- Algorithms for demographic inference and personality profiling
- Anonymization of demographic and personal traits

## Part 5: Democracy and the language of manipulation

- Approaches to identify propaganda and manipulation in news
- Fake news
- Political and media framing

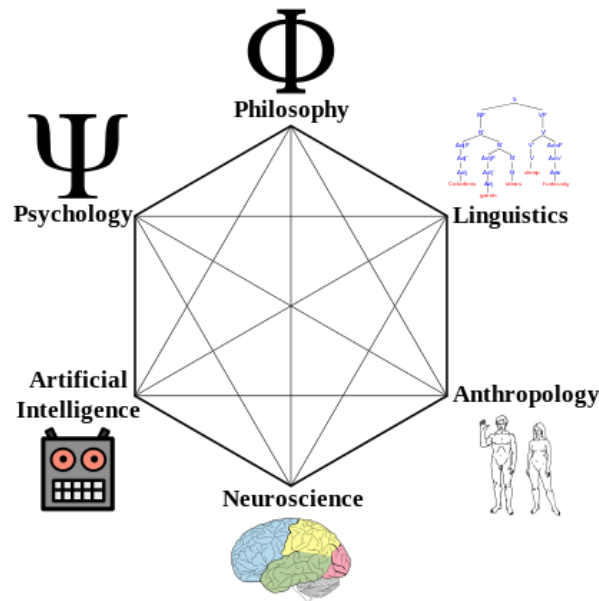
## Part 6: NLP for Social Good

- Low-resource NLP, applications for disaster response and monitoring diseases
- Medical applications, psychological counseling
- Interfaces for accessibility.



# Ethics and NLP are Interdisciplinary

- Philosophy
- Sociology
- Psychology
- Linguistics
- Sociolinguistics
- Social psychology
- Computational Social Science
- Machine Learning



# 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

# Quiz break!

Switch to menti.com

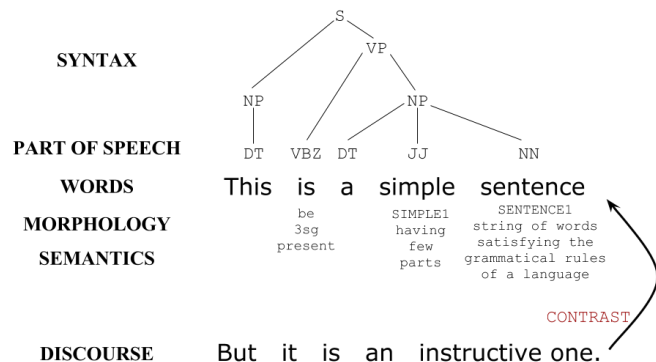
# Outline

**UKP Lab: profile and projects**

**Administrative course issues**

**Introduction to Ethics in NLP**

# What NLP Has To Do With Ethics?



## Applications

- Machine Translation
- Information Retrieval
- 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
- ...

# 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.

Herbert H. Clark & Michael F. Schober, 1992

Decisions we make about our data, methods, and tools are tied up with their impact on people and societies.

# What is Ethics?

“Ethics is a study of what are **good and bad** ends to pursue in life and what it is **right and wrong** to do in the conduct of life.

It is therefore, above all, a **practical discipline**.

Its primary aim is to determine how one ought to live and what actions one ought to do in the conduct of one's life.”

-- Introduction to Ethics, John Deigh

# What is Ethics?

It's the **good** things

It's the **right** things



# What is Ethics?

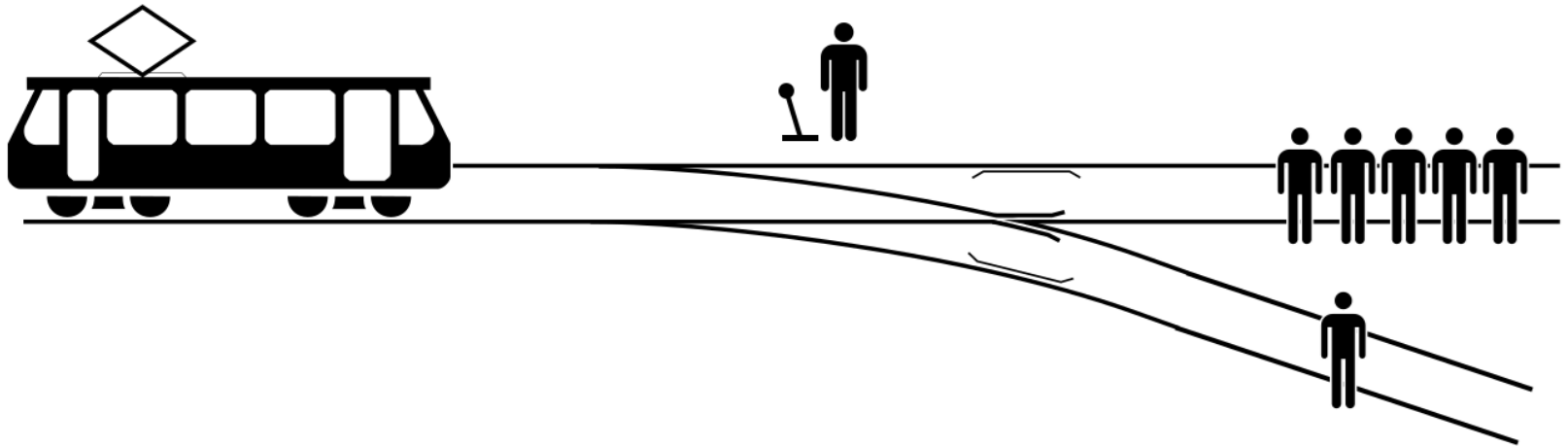
It's the **good** things

It's the **right** things

How simple is it to define  
what's good and what's right?

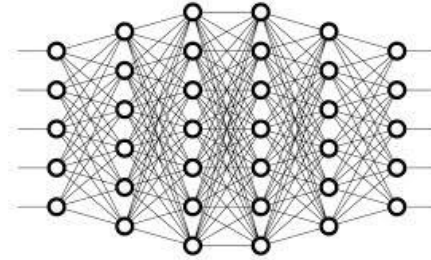
# The Trolley Dilemma

Should you pull the lever to divert the trolley?



[From Wikipedia]

# Let's Train a Chicken Classifier



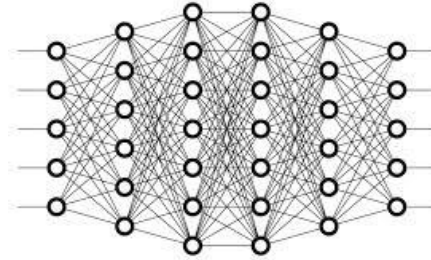
rooster



hen



# Let's Train a Chicken Classifier



rooster

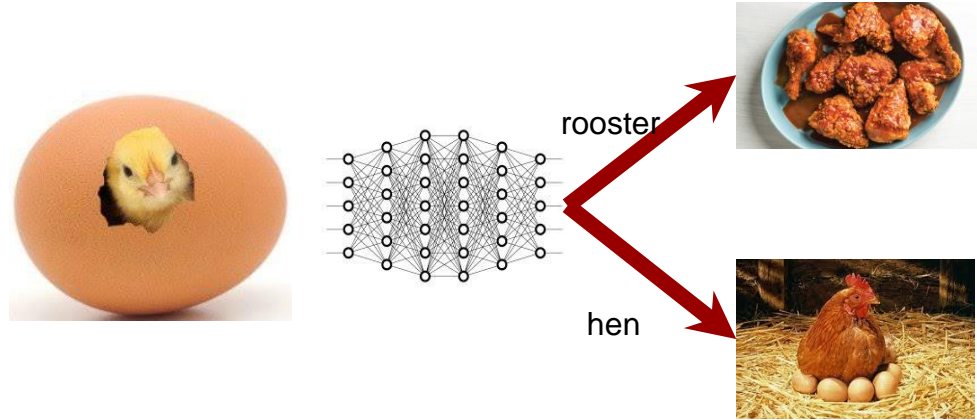


hen



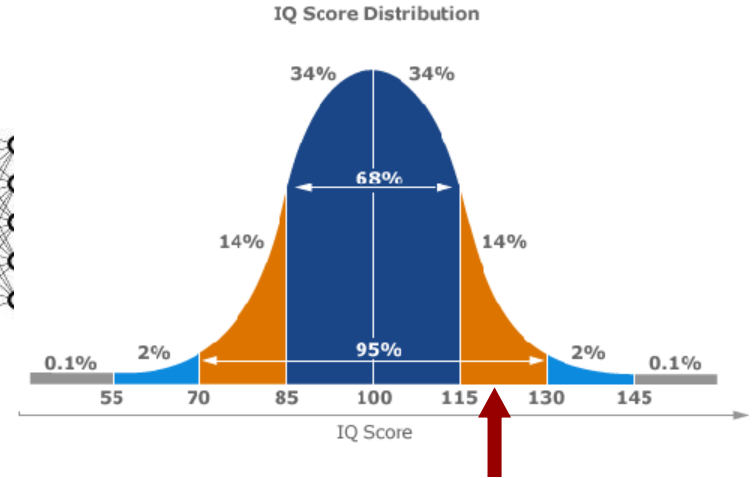
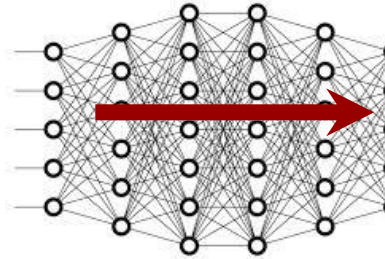
## Ethical?

# Let's Train a Chicken Classifier



- Ethics is inner guiding, moral principles, and values of people and society
- Ethics is not black and white; there are grey areas.  
This course will not give binary answers.
- Ethics changes over time with values and beliefs of people
- Legal  $\neq$  ethical

# Let's Train an IQ Classifier



- **Intelligence Quotient:** a number used to express the apparent relative intelligence of a person

# An IQ Classifier

Let's train a classifier to predict people's IQ from their photos.

- Who could benefit from such a classifier?

# An IQ Classifier

Let's train a classifier to predict people's IQ from their photos.

- Who could benefit from such a classifier?
- Assume the classifier is 100% accurate. Who can be harmed from such a classifier?



# An IQ Classifier

Let's train a classifier to predict people's IQ from their photos.

- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Our test results show 90% accuracy
  - We found out that white females have 95% accuracy
  - People with blond hair under age of 25 have only 60% accuracy

# An IQ Classifier

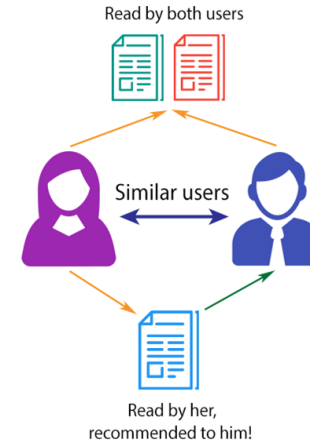
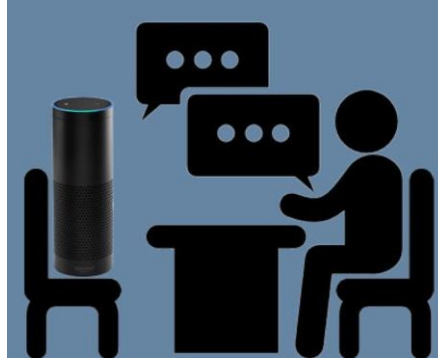
Let's train a classifier to predict people's IQ from their photos.

- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Our test results show 90% accuracy
  - We found out that white females have 95% accuracy
  - People with blond hair under age of 25 have only 60% accuracy
- Who is responsible?
  - Researcher/developer? Reviewer? University? Society?

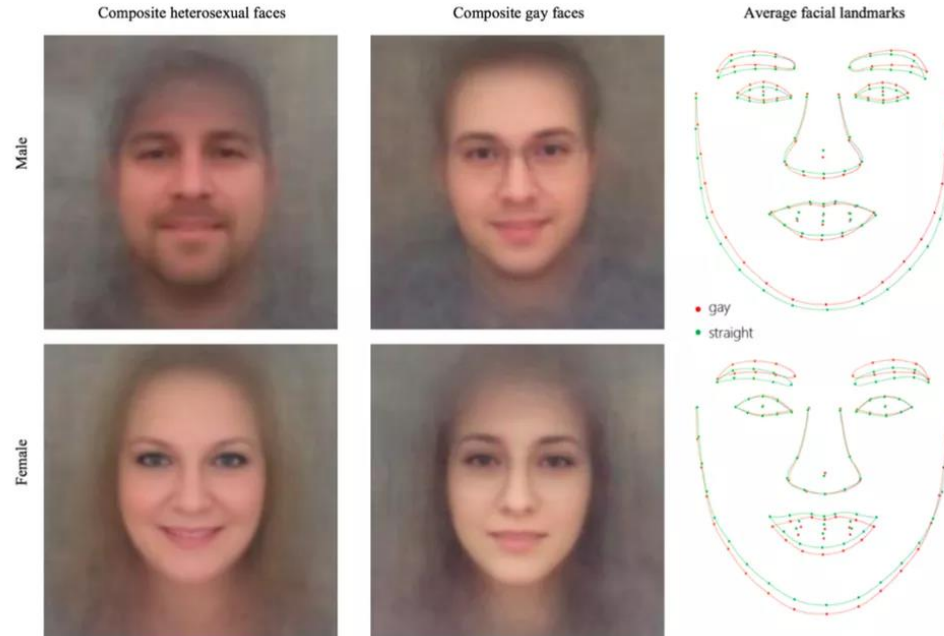
# What's the Difference?



# AI and People



# A Case Study: the “A.I. Gaydar”



# A Case Study: the “A.I. Gaydar”

**Abstract.** We show that faces contain much more information about sexual orientation than can be perceived and interpreted by the human brain. We used deep neural networks to extract features from 35,326 facial images. These features were entered into a logistic regression aimed at classifying sexual orientation. Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 81% of cases, and in 74% of cases for women. Human judges achieved much lower accuracy: 61% for men and 54% for women. The accuracy of the algorithm increased to 91% and 83%, respectively, given five facial images per person. Facial features employed by the classifier included both fixed (e.g., nose shape) and transient facial features (e.g., grooming style). Consistent with the prenatal hormone theory of sexual orientation, gay men and women tended to have gender-atypical facial morphology, expression, and grooming styles. Prediction models aimed at gender alone allowed for detecting gay males with 57% accuracy and gay females with 58% accuracy. Those findings advance our understanding of the origins of sexual orientation and the limits of human perception. Additionally, given that companies and governments are increasingly using computer vision algorithms to detect people’s intimate traits, our findings expose a threat to the privacy and safety of gay men and women.

Wang & Kosinski. **Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.** *Journal of Personality and Social Psychology* (in press). September 7,

# A Case Study: the “A.I. Gaydar”

- Research question
  - Identification of sexual orientation from facial features
- Data collection
  - Photos downloaded from a popular American dating website
  - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly
- Method
  - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
  - 81% for men, 74% for women

# Let's Discuss...

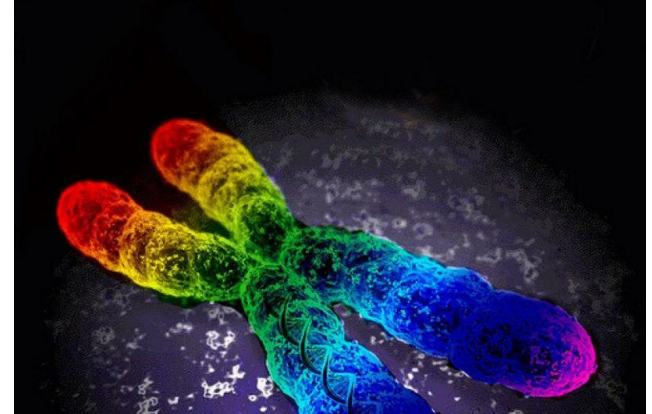
- Research question
  - Identification of sexual orientation from facial features
- Data collection
  - Photos downloaded from a popular American dating website
  - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented equally
- Method
  - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
  - 81% for men, 74% for women

**Is anything wrong here?**



# Let's Discuss...

- Research question
  - Identification of sexual orientation from facial features



- Identification of sexual orientation from facial features

## How people can be harmed by this research?

- In many countries being gay person is prosecutable (by law or by society)
- Might affect people's employment, family relationships
- Personal attributes, e.g. gender, race, sexual orientation, religion are social constructs. They can change over time. They can be non-binary. They are private, intimate, often not visible publicly.
- Importantly, these are properties for which people are often discriminated against.

# Research Question

*“... Additionally, given that companies and governments are increasingly using computer vision algorithms to detect people’s intimate traits, our findings expose a threat to the privacy and safety of gay men and women.”*

→ your thoughts on this?

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly



- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

**Legal  $\neq$  Ethical**

**Public  $\neq$  Publicized**

**Did these people agree to participate in the study?**

**→ Violation of social contract**

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly



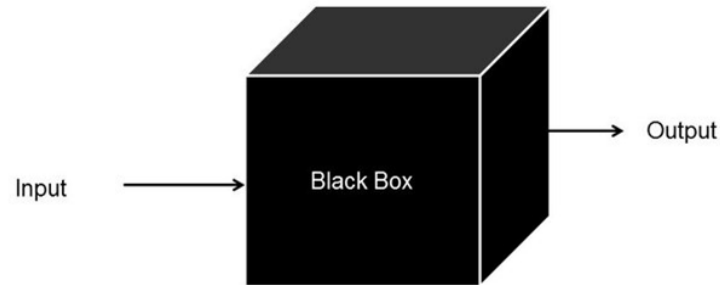
- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

**Only white people, who self-disclose their orientation, certain social groups, certain age groups, certain time range/fashion;**

**the photos were carefully selected by subjects to be attractive so there is even self-selection bias...**

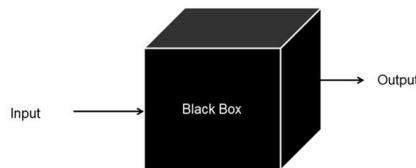
**The dataset is balanced, which does not represent true class distribution.**

- A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification





- A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification



- **can we use not interpretable models when we make predictions about sensitive attributes, about complex experimental conditions that require broader world knowledge?**
- **how to deal with bias amplification?**

# Evaluation

- Accuracy: 81% for men, 74% for women

# The Cost of Misclassification



# The Cost of Misclassification



# Dual Use Problem



*“We live in a dangerous world, where harm doers and criminals easily mingle with the general population; the vast majority of them are unknown to the authorities.*

*As a result, it is becoming ever more challenging to detect anonymous threats in public places such as airports, train stations, government and public buildings and border control. Public Safety agencies, city police department, smart city service providers and other law enforcement entities are increasingly strive for Predictive Screening solutions, that can monitor, prevent, and forecast criminal events and public disorder without direct investigation or innocent people interrogations. “*

# The Dual Use of A.I. Technologies

- Who should be responsible?
  - The person who uses the technology?
  - The researcher/developer?
  - Paper reviewers?
  - University?
  - Society as a whole?

**We need to be aware of real-world impact of our research and understand the relationship between ideas and consequences**

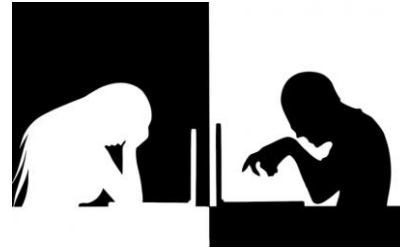
# Learn to Assess AI Systems Adversarially



TECHNISCHE  
UNIVERSITÄT  
DARMSTADT

- Who could benefit from such a technology?
- Who can be harmed by such a technology?
- Representativeness of training data
- Privacy Issues
- Could sharing this data have major effect on people's lives?
- Does the system optimize for the “right” objective?
- What are confounding variables and corner cases to control for?
- Could prediction errors have major effect on people's lives?

# We Cannot Foresee All Possible Uses of Technology





# “Smart” Toys: Barbie

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

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

“Yeah,” Tiara said. “I only have one.”

“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.



# Watch This Talk

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.

<https://goo.gl/8tBho8>

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.



Blog posts:

<https://psmag.com/economics/is-the-trolley-problem-derailing-the-ethics-of-self-driving-cars>

<https://medium.com/@yonatanzunger/asking-the-right-questions-about-ai-7ed2d9820c48>

# Next Lecture

## (Theoretical and Philosophical) Foundations Part II