

Course

Foundations of Artificial Intelligence

Lectures 12 and 13 – Ethics and Wrap up

Dr. Mohsen Mesgar

Universität Duisburg-Essen

Ethics for AI

Motivation

- AI increasingly has an impact on everything from social media to healthcare.
- AI is used to make credit card decisions, to conduct video surveillance in airports, and to inform military operations.
- if it is true that AI solves old problems, it is also true that it raises new ones.
- These technologies have the potential to harm or help the people that they serve.
- By applying an ethical lens, we can work toward identifying the harms that these technologies can cause to people and we can design and build them to reduce these harms - or decide not to build them.

Motivation

- How do automated decision systems powered by AI affect human decisions?
- Why is privacy important not only as individual good but also as a common one?
- If autonomous systems, either weapons or vehicles, are designed to decide about the life and death of people, what about human dignity?
- What is the moral responsibility of AI designers in order not only to avoid negative effects of technologies, but also to promote positive ones?

Motivation

- What is the impact of these tools on the notion of governance and what are the effects of algorithmic governance on individuals and societies?
- How fairness, equality and justice should be included in these AI systems and how should they be managed by them?

What is Ethics?

Ethics is difficult to define: in a nutshell we can say that it is the systematic reflection on what is moral.

Branch of Philosophy

Ethics is the **philosophical study of morality**. It is the study of what are good and bad ends to pursue in life and what is **right** and **wrong** to do in the conduct of life. It is [...] primarily a **practical** discipline.

(Deigh, 2010, p. 7)

Synonym for Moral Code

Sometimes “ethics” is used to refer to the **moral code** or system of a particular tradition.

Examples: Christian ethics,
professional ethics

What is Ethics?

- What is important to remember is that **ethics is not a set of rules**.
- Ethics is not **a manual with answers**, nor **a checklist**.
- Rather ethics is **a process**, and a complex one indeed.
- It reflects on **questions and arguments concerning the moral choices people make** and, for this reason, it cannot be determined once and for all.

What is Morality?

Universal Concept

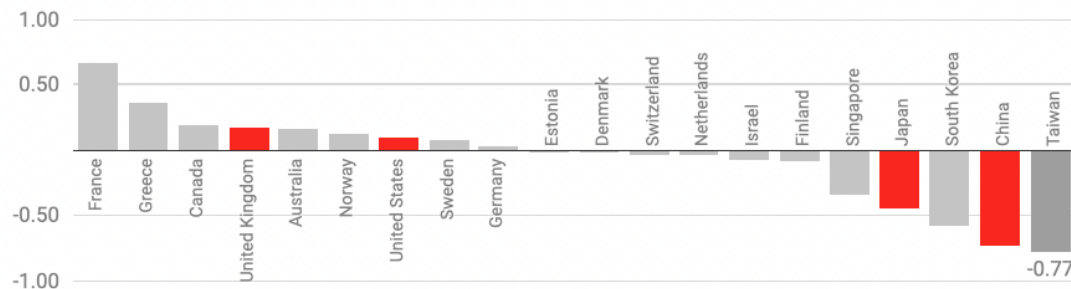
Universal ideal of what one ought to do or ought not to do, guided by reason / rational grounds.

Conventional System of Community

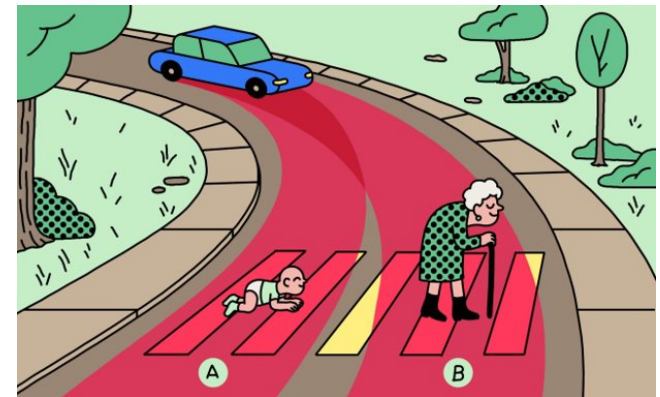
The members' shared beliefs about wrong and right, good and evil, and the corresponding customs and practices that prevail in the society.

Whose Life Matters More?

Countries with more individualistic cultures are more likely to spare the young



A comparison of countries piloting self-driving cars: If the bar is closer to 1, respondents placed a greater emphasis on sparing the young; if the bar is closer to -1, respondents placed a greater emphasis on sparing the old; 0 is the global average.



Try it out!

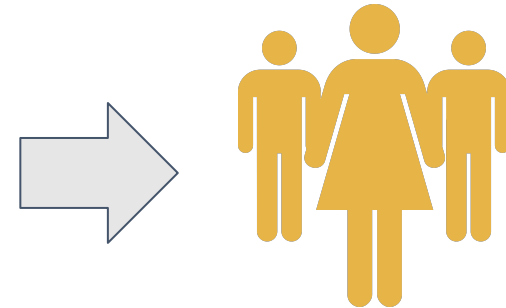
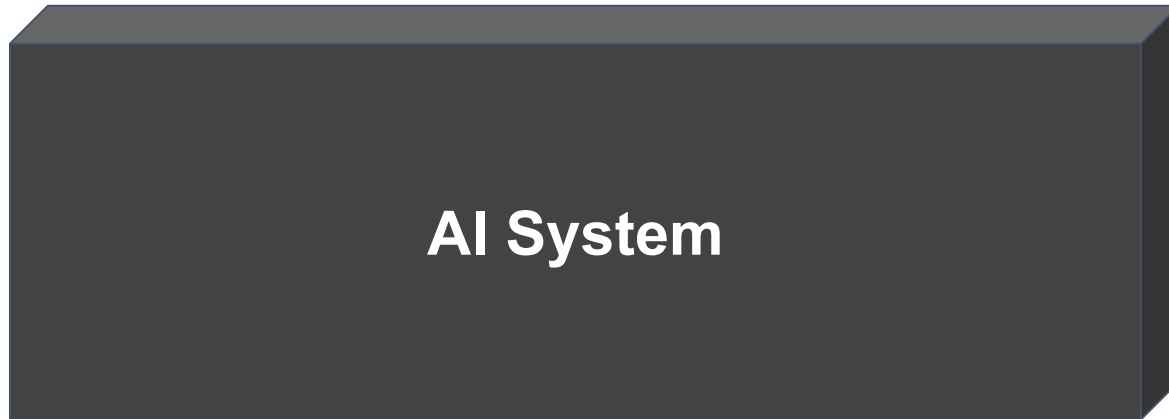
<http://moralmachine.mit.edu/hl/de>

Should we first vaccinate vulnerable older people who are more isolated, or younger people who are more social?

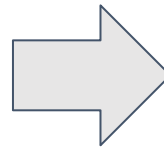
Moral vs. Legal

	legal	illegal
moral	Doing your homework	Civil disobedience
immoral	Cheating on your spouse	Murder

Source of Harm - Direct



analyzing medical
documents



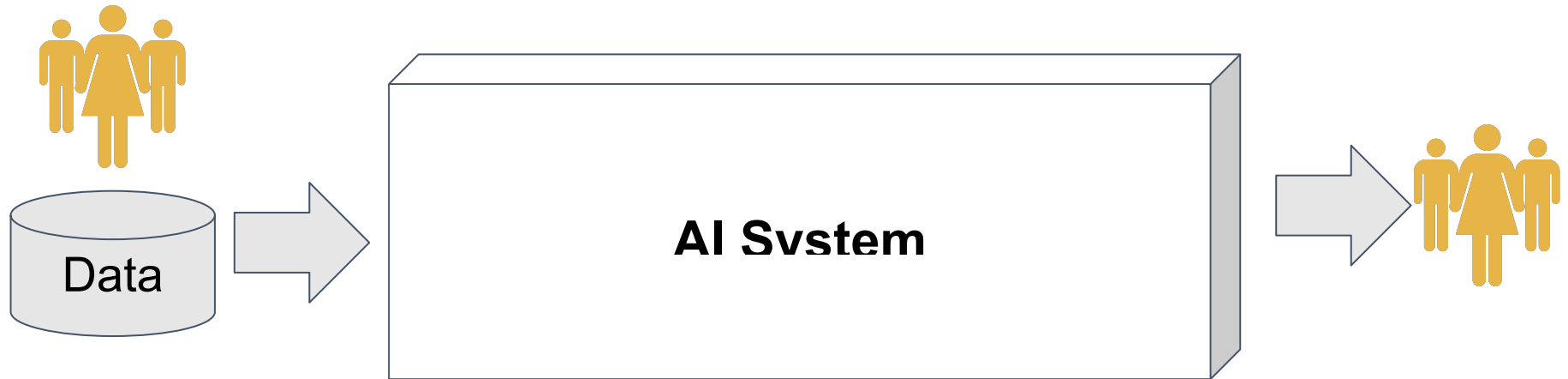
drug overdose
killing the patient

Dual Use

AI Task	Beneficial Use	Malicious Use
Hate speech detection	Fighting hate crimes	Censorship of free speech
Detection of fake news / reviews	Fighting misinformation	Generation of fake news / reviews
...

Can you think of other AI tasks that have beneficial but also potentially malicious uses?

Source of Harm - Bias



Doctor vs. Nurse

The doctor recommended to perform an X-ray.

He/She said ...



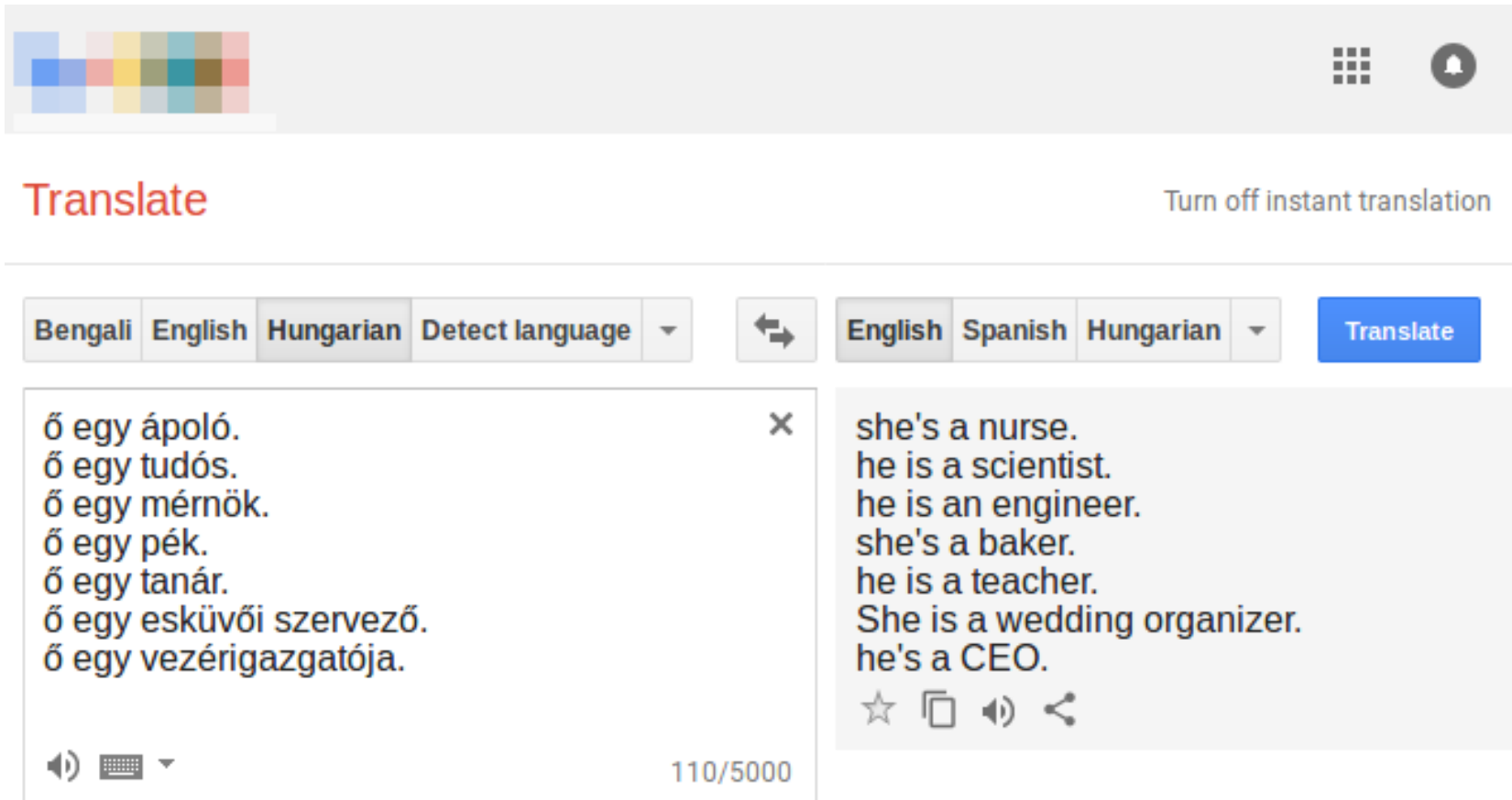
The nurse recommended to perform an X-ray.

He/She said ...

Do you think “he” or “she” is a more likely continuation in the above cases (respectively)?

What would happen if you asked an AI system (with language capabilities)?

Bias in Machine Translation



The image shows a screenshot of the Google Translate web interface. At the top, there's a header with the University of Duisburg-Essen logo and the motto "Offen im Denken". Below the header, the word "Translate" is prominently displayed in red. To the right of "Translate" is a link that says "Turn off instant translation". Below this, there are two language selection bars. The left bar has "Bengali", "English", "Hungarian", and "Detect language" with a dropdown arrow. The right bar has "English", "Spanish", and "Hungarian" with a dropdown arrow. A blue "Translate" button is to the right of the second bar. Below the language bars, there are two text boxes. The left box contains Hungarian text: "ő egy ápoló.", "ő egy tudós.", "ő egy mérnök.", "ő egy pék.", "ő egy tanár.", "ő egy esküvői szervező.", "ő egy vezérigazgatója.". The right box contains the English translation: "she's a nurse.", "he is a scientist.", "he is an engineer.", "she's a baker.", "he is a teacher.", "She is a wedding organizer.", "he's a CEO.". At the bottom of the left box, there are icons for a speaker and a keyboard, and a character count "110/5000".

Translate [Turn off instant translation](#)

Bengali English Hungarian Detect language ↕ English Spanish Hungarian Translate

ő egy ápoló.
ő egy tudós.
ő egy mérnök.
ő egy pék.
ő egy tanár.
ő egy esküvői szervező.
ő egy vezérigazgatója.

she's a nurse.
he is a scientist.
he is an engineer.
she's a baker.
he is a teacher.
She is a wedding organizer.
he's a CEO.

110/5000

What is Bias?

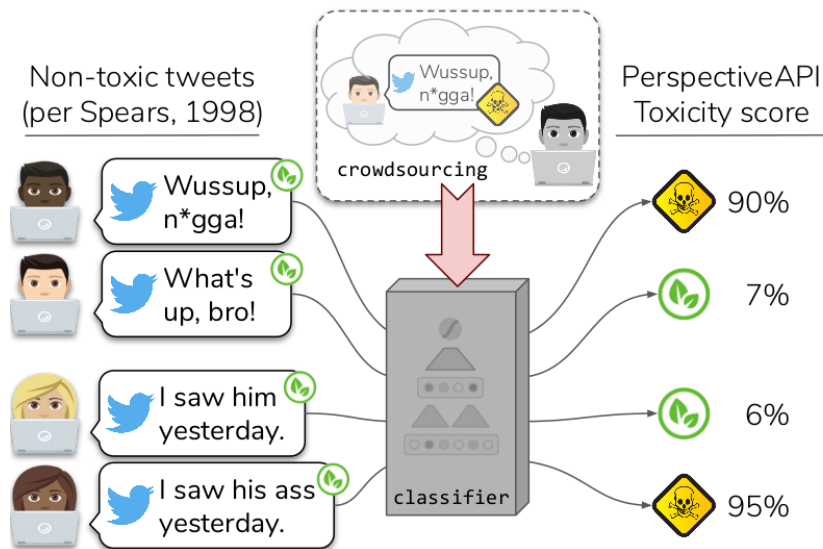
- **Cognitive bias** arises due to the tendency of the human mind to categorize the world.
→ simplifies processing.

Social biases in data, algorithms, and applications

Statistical bias in machine learning

- **Inductive bias:** assumptions made by model about target function to generalize from data

Why is Bias Problematic? (Social View)



AI Applications

Employment matching, advertisement placement, parole decisions, search, chatbots, face recognition, ...

Social Stereotypes

Gender, Race, Disability, Age, Sexual orientation, Culture, Class, Poverty, Language, Religion, National origin, ...

Why is Bias Problematic?



COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN



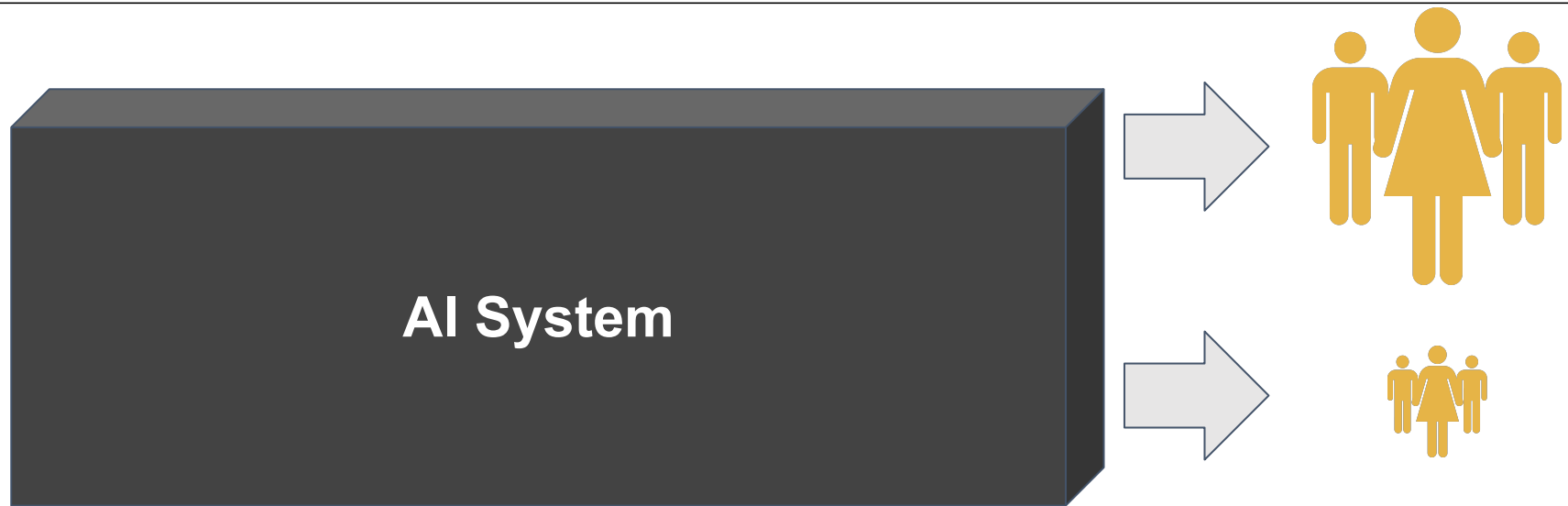
COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE



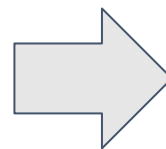
COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

Because a “COOKING” event is taking place, the model is more likely to predict the agent to be a woman.
(Zhao et al., 2017)

Source of Harm - Unfair Outcomes



filtering
job applications



Better chances for
people living in a
certain area

Fairness

- Treating everyone equally is fair, right?
- So, everyone gets the same grade from now on
-

fundamental principle of justice

“equals should be treated equally and
unequals unequally”



Bild von [Gordon Johnson](#) auf [Pixabay](#)

Group vs. Individual Fairness

group fairness

- errors should be distributed similarly across protected groups

Which groups are/should be protected?

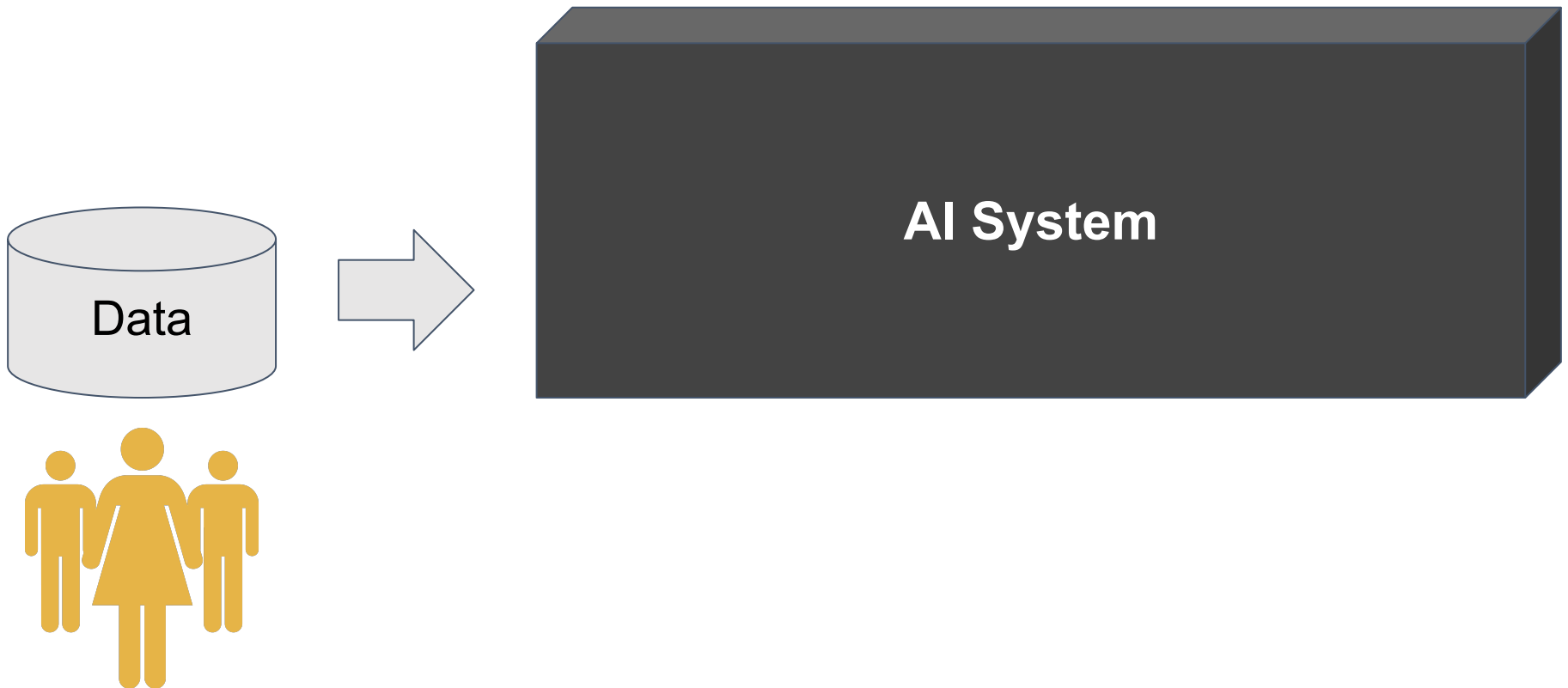
individual fairness

- similar individuals should be treated similarly regardless of group membership

How can we measure similarity of individuals?

cannot reach group and individual fairness at the same time

Source of Harm - Input/Training Data



Privacy

“I’ve got nothing to hide.”



Do you have curtains? / Do you close your shutters at night?
Can I see your credit card bills from last year?

A Taxonomy of Privacy (Solove, 2007)

Problems and harms related to privacy

Privacy = intimacy?

Privacy = the right to be let alone?

“Privacy [...] is a plurality of different things that do not share one element in common but that nevertheless bear a resemblance to each other.”

Information Collection

Surveillance

Interrogation

Information Processing

Aggregation

Identification

Insecurity

Secondary Use

Exclusion

Information Dissemination

Breach of Confidentiality

Disclosure

Exposure

Increased Accessibility

Blackmail

Appropriation

Distortion

Invasion

Intrusion

Decisional Interference

Data Privacy Regulations

European Regulation 2016/679
General Data Protection Regulation ([GDPR](#))



Main rights of the “data subject” (natural person):

- Right of access
- Right of rectification
- Right to erasure (“right to be forgotten”)
- Right to withdraw consent at any time
- Right to lodge a complaint with a supervisory authority
- Right to restriction of processing
- Right to data portability

Similar laws in the US:
California Consumer Privacy Act

Data Privacy vs. Data Ethics

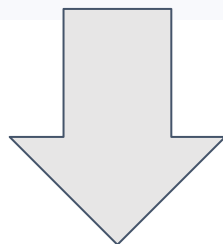
- Data privacy is responsibly collecting, using and storing data about people, in line with the expectations of those people, customers, regulations and laws
- Data ethics is doing the right thing with data, considering the human impact from all sides, and making decisions based on your values.

[based on: [Lawler, 2019](#)]

“Just because we can do something, doesn’t mean we should.”
[Should a company sell user information to political campaigns?](#)

Anonymization (De-Identification)

Informe clínico del paciente : Paciente **varón** de **70** años de edad ,
minero jubilado , sin alergias medicamentosas conocidas . Operado de
una hernia el **12** de **enero** de **2016** en el **Hospital Costa del**
Sol por la Dra . **Juana López** . Derivado a este centro el día 16 del
mismo mes para revisión .



Informe clínico del paciente : Paciente **SEX** de **AGE** **AGE** de edad ,
PROFESSION jubilado , sin alergias medicamentosas conocidas .
Operado de una hernia el **DATE** **DATE** **DATE** **DATE** **DATE** en el
HOSPITAL **HOSPITAL** **HOSPITAL** **HOSPITAL** por la Dra .
DOCTOR **DOCTOR** . Derivado a este centro el día 16 del mismo mes
para revisión .

HitzalMed
(Lopez et al., 2020)

After having run some
anonymization system
on our data, is
everything fine?

Literature – Ethics in AI

Overviews

- Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use Ai in a responsible way*. Cham, Switzerland: Springer.
- Fort, K., & Couillault, A. (2016). Yes, We Care! Results of the Ethics and Natural Language Processing Surveys. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Retrieved from <https://www.aclweb.org/anthology/L16-1252>
- Hovy, D., & Spruit, S. L. (2016). The Social Impact of Natural Language Processing. In K. Erk & N. A. Smith (Chairs), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany. Retrieved from <https://www.aclweb.org/anthology/P16-2096.pdf>

Overviews

- Leidner, J. L., & Plachouras, V. (2017). Ethical by Design: Ethics Best Practices for Natural Language Processing. In D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube, & H. Wallach (Chairs), Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. Retrieved from <https://www.aclweb.org/anthology/W17-1604.pdf>
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 205395171667967. Retrieved from <https://journals.sagepub.com/doi/pdf/10.1177/2053951716679679>
- Zweig, K. A. (2019). *Ein Algorithmus hat kein Taktgefühl: wo Künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können*. München: Heyne.

Bias

- Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions. In D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube, & H. Wallach (Chairs), Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. Retrieved from <https://www.aclweb.org/anthology/W17-1606.pdf>
- Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2019). Assessing gender bias in machine translation: A case study with Google Translate. Neural Computing and Applications, 14(1). Retrieved from <https://arxiv.org/pdf/1809.02208.pdf>
- Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/P19-1164.pdf>

Bias

- Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. In Proceedings of the Third Workshop on Abusive Language Online (pp. 25–35). Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W19-3504.pdf>
- Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1668–1678). Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/P19-1163.pdf>

Literature – Ethics in AI

Fairness

- Loukina, A., Madhani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 1–10). Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W19-4401.pdf>

Literature – Ethics in AI

Gender Stereotypes

- Bhaskaran, J., & Bhallamudi, I. (2019). Good Secretaries, Bad Truck Drivers? Occupational Gender Stereotypes in Sentiment Analysis. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing (pp. 62–68). Association for Computational Linguistics. Retrieved from <https://www.aclweb.org/anthology/W19-3809.pdf>

Thank You

Title

Title

Title

Title

Title

Title

Title

Title

Title

Title

Title

Title

Title

Title

Reading

Mandatory:

-