

Ethical Aspects of Machine Learning (and AI)

Winter school „**CAS ADS M6 Deep Learning**“

Hotel Regina, Mürren

26.1.2022

Claus Beisbart

Claus.Beisbart@philo.unibe.ch

Aims of this lecture:

Raise awareness of ethical challenges
related to machine learning

Discuss some challenges using
methods and concepts from ethics

Aims of this ~~lecture~~ conversation

Raise awareness of ethical challenges
related to machine learning

Discuss some challenges using
methods and concepts from ethics

Method

1. Start from **your** experience, views
2. Give input from **philosophy**
3. Discuss

Steps

1. Collect issues

2. Discuss one or two of them

Question for you

You are attending a winter school on ML. Using the competences that you get here, for which purposes do you want to use ML (in your job, studies ...)?

Question for you

What benefits (for society, humankind, ...) do such applications of ML promise in **your** view?

Philosophical input

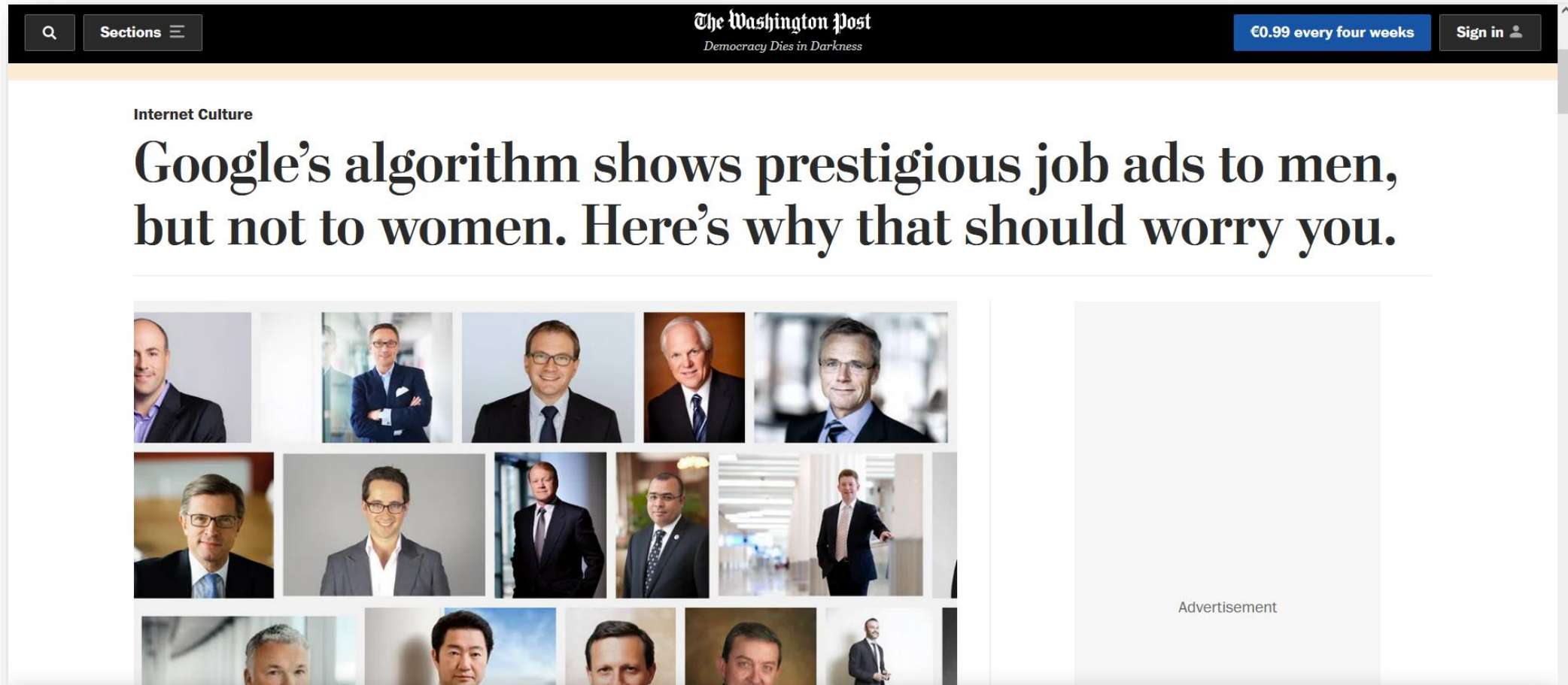
What benefits (for society, humankind, ...) do such applications of ML promise in your view?

Due to the condition in brackets, I was asking for an ethical evaluation.

Question for you

To what harms, risks or challenges (for society, humankind, ...) may such ML applications lead in **your** view?

Example 1



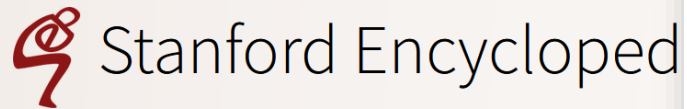
Example 2



Figure 1. (Left) Location of the crash on northbound Mill Avenue, showing the paths of the pedestrian in orange and of the Uber test vehicle in green. (Right) Postcrash view of the Uber test vehicle, showing damage to the right front side.

Philosophical input: overview of debates

- 2. Main Debates
 - 2.1 Privacy & Surveillance
 - 2.2 Manipulation of Behaviour
 - 2.3 Opacity of AI Systems
 - 2.4 Bias in Decision Systems
 - 2.5 Human-Robot Interaction
 - 2.6 Automation and Employment
 - 2.7 Autonomous Systems
 - 2.8 Machine Ethics
 - 2.9 Artificial Moral Agents
 - 2.10 Singularity



Browse About Support SEP

Entry Contents

Bibliography

Academic Tools

Friends PDF Preview



Author and Citation Info



Back to Top



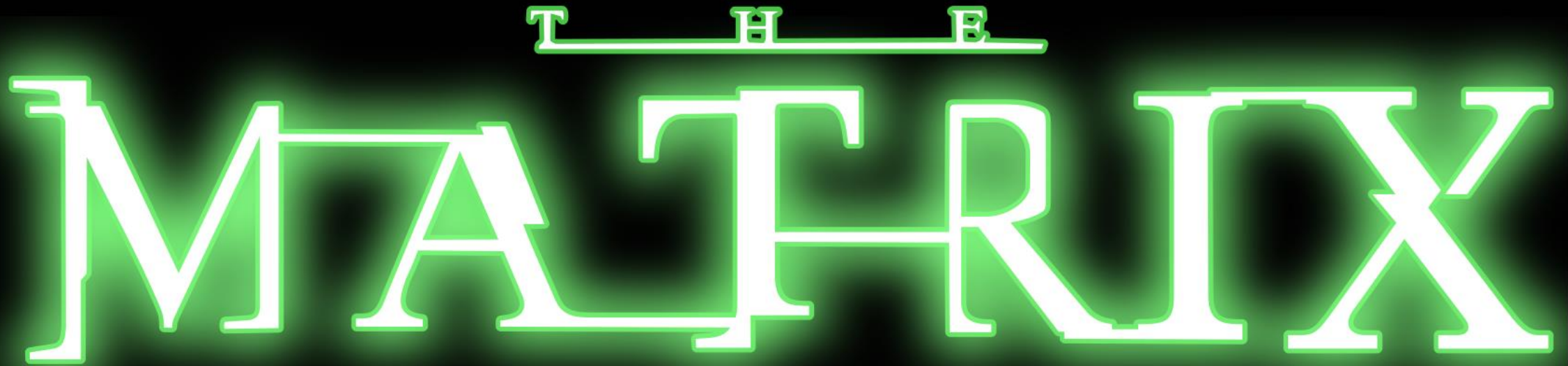
Ethics of Robotics

First published Th

Artificial intelligence (AI) and robotics are digital technologies that will have significant impact on the development of humanity in the near future. They have raised fundamental questions about what we should do with these systems, what the systems themselves should do, what risks they involve, and how we can control these.

Müller, V. C., [Ethics of Artificial Intelligence and Robotics](#), *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.).

Getting started

The Matrix logo is displayed on a black background. It features the word "MATRIX" in a large, white, serif font with a bright green glow. Above the word, the words "THE" are written in a smaller, white, serif font, also with a green glow, and are underlined by a horizontal green line.

„The Matrix“



Autonomy

2010

«Eine einzige Simulation genügte, um in die Schicksale von Millionen von Menschen einzugreifen und Europa lahmzulegen.

[...]

Plötzlich werden alle zu Zuschauern: die Fluggäste, die Piloten, die Airlines, der Wetterdienst, die Behörden. Die „human response“, die menschliche Antwort auf die Maschine, ist nicht mehr möglich, weil auch in den menschlichen Entscheidungsgruppen ein Programm von Befehlen, Verordnungen und Routinen abläuft.»

Frank Schirrmacher

Árni Friðriksson, wikimedia commons ([CC BY-SA 3.0](https://commons.wikimedia.org/wiki/File:2010-04-19_01.jpg)); F.A.Z., 19.4.2010



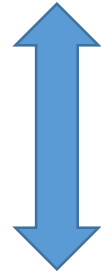
Does AI pose a threat
to human autonomy?

Philosophical input: concept

Gr. autos: self

Gr. nomos: law

autonomy
Self determination



Heteronomy
Being determined by others

Philosophical input: Immanuel Kant



(1724–1804)

„Autonomie des Willens ist die Beschaffenheit des Willens, dadurch derselbe ihm selbst (unabhängig von aller Beschaffenheit der Gegenstände des Wollens) ein Gesetz ist.“

Grundlegung zur Metaphysik der Sitten, Akademie-Ausgabe IV, 440

Philosophical input: Kant's ethics of autonomy



(1724–1804)

„Autonomie ist also der Grund
der Würde der menschlichen und
jeder vernünftigen Natur.“



„Die Handlung, die mit der
Autonomie des Willens
zusammen bestehen kann, ist
erlaubt; die nicht damit stimmt,
ist unerlaubt.“

Philosophical input: ethics of medicine

Respect for autonomy:

1. „Tell the truth.
2. Respect the privacy of others.
3. Protect confidential information.
4. Obtain consent for interventions with patients.
5. When asked, help others to make decisions.“

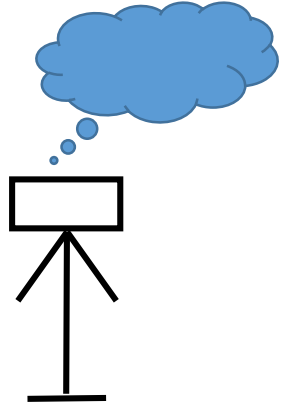
T. L. Beauchamp & J. F. Childress, Principles of Biomedical Ethics, New York 2001⁵, 65



Question for you

How do you think may ML/AI
applications impact on human
autonomy?

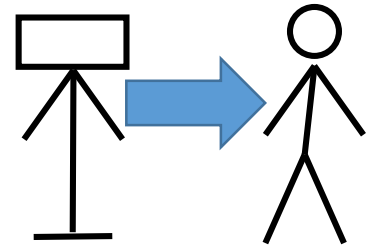
Philosophical input: 2 routes to problems



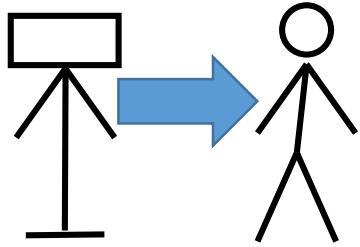
AI applications take decisions

Where is the problem?

AI applications influence human decisions



Philosophical input: concepts



Sorts of influence:

- Conditioning
- Nudging
- Herding

Cf.. Köszegi, S. T., Der autonome Mensch im Zeitalter des Digitalen Wandels, in: Hengstschläger, Markus/Rat für Forschung und Technologieentwicklung (Hrsg.), Digitaler Wandel und Ethik, Wien 2020, 62-86

Der autonome Mensch im Zeitalter des Digitalen Wandels

Sabine Theresia Köszegi

Einleitung

Seit etwa 50 Jahren setzen Menschen modell- und datenbasierte Unterstützungssysteme zur Entscheidungsfindung – vorwiegend bei komplexen Entscheidungsproblemen – ein. Menschen ha-

Example: nudging

The screenshot shows a web browser window with the URL <https://www.tagesanzeiger.ch>. The browser's tab bar includes several open tabs: 'Claus' Links', 'Ethics of /', 'Neuer Tab', 'Startpage', 'Questioni', 'Google's', 'heteron', 'fremdbest', 'Principles', 'The New', and 'Tages-/X'. The browser's address bar shows the URL and a zoom level of 133%. The website's header features the 'TagesAnzeiger' logo and navigation links like 'Zürich' and 'Meinung'. A large white pop-up is centered on the screen, featuring an illustration of a man in a red jacket and glasses running towards a yellow square. Below the illustration, the text reads: 'Jetzt alle Artikel 14 Tage kostenlos lesen'. At the bottom of the pop-up, there is a blue button labeled 'OK' and a link to 'Privatsphäre-Einstellungen'. Below the pop-up, a cookie consent banner is visible, stating: 'Wir benutzen Cookies und andere Technologien'. The banner includes a paragraph explaining the use of cookies and a link to the 'Datenschutzerklärung'.

Jetzt alle Artikel 14 Tage kostenlos lesen

Mit Ihrer E-Mail-Adresse registrieren (keine Kreditkarte nötig) und 14 Tage

Wir benutzen Cookies und andere Technologien

Diese Webseite verwendet Cookies und andere Technologien, um Ihre Seitennutzung auszuwerten und Ihnen nutzungsbasiert redaktionelle Inhalte und Werbung anzuzeigen. Das ist für uns wichtig, denn unser Angebot finanziert sich über Werbung. Durch das Klicken auf OK oder durch die Nutzung der Website stimmen Sie diesen Datenbearbeitungen zu. Weitere Informationen, wie Sie z.B. Ihre Zustimmung jederzeit widerrufen können, finden Sie unter den Privatsphäre-Einstellungen ihres Browsers sowie in der [Datenschutzerklärung](#).

OK

[Privatsphäre-Einstellungen](#)

Philosophical input: concept

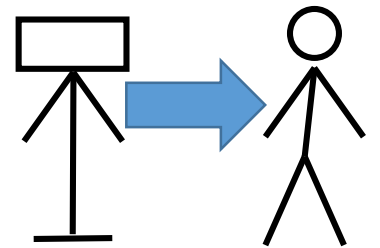
Not every influence on human decisions is bad.

One problem:

Manipulation

„(disapproving) behaviour that controls or influences somebody/something, often in a dishonest way so that they do not realize it“

[Oxford Learner's Dictionaries](#)



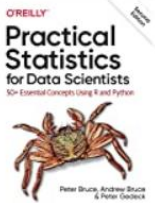
Example: recommender system

Claus' Links Ethics of Artificial Intelligence Neuer Tab Startpage Sucher Questioning the Google's algorithm heteronomy - Di fremdbestimmung ILIAS Universität Amazon.com: X + - □ X

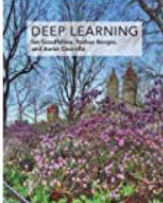
← → ↻ https://www.amazon.com/-/de/dp/1492032646/ref=sr_1_1?_mk_de_DE=ÄMÄZÖN&crd=35G6F2FAIO089&keywords=machine+learning&qid=1643116789&spri ☆ ⌵ ☰

Kunden, die diesen Artikel angesehen haben, haben auch angesehen


Seite 1 von 8




Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python
Peter Bruce
★★★★☆ 343
Taschenbuch
25,00 \$
Erhalten Sie es bis Montag, 7. Februar
11,04 \$ Versand




Deep Learning (Adaptive Computation and Machine Learning series)
> Ian Goodfellow
★★★★☆ 1.558
Gebundene Ausgabe
43,00 \$
Erhalten Sie es bis Montag, 7. Februar
15,02 \$ Versand




Introduction to Machine Learning with Python: A Guide for Data Scientists
Andreas C. Müller
★★★★☆ 440
Taschenbuch
39 Angebote ab 30,35 \$



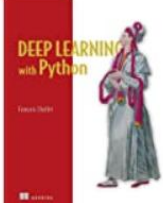
Python Data Science Handbook: Essential Tools for Working with...
> Jake VanderPlas
★★★★☆ 474
Taschenbuch
37,50 \$
Erhalten Sie es bis Montag, 7. Februar
12,52 \$ Versand



The Hundred-Page Machine Learning Book
> Andriy Burkov
★★★★★ 728
Taschenbuch
39,95 \$
9,80 \$ Versand



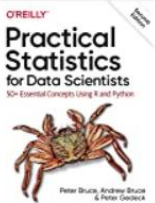
Python Machine Learning: Machine Learning and Deep Learning with Python, ...
> Sebastian Raschka
★★★★☆ 300
Taschenbuch
22 Angebote ab 29,90 \$



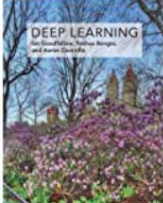
Deep Learning with Python
> Francois Chollet
★★★★★ 1.053
Taschenbuch
Bestseller Nr. 1 in Sprach- & Audioverarbeitung
16,50 \$
Erhalten Sie es bis Montag, 7. Februar
12,07 \$ Versand

Kunden, die diesen Artikel gekauft haben, kauften auch


Seite 1 von 13




Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python
Peter Bruce, Andrew Bruce & Peter Goodrich




Deep Learning (Adaptive Computation and Machine Learning series)
> Ian Goodfellow




Python Data Science Handbook: Essential Tools for Working with...
> Jake VanderPlas




Introduction to Machine Learning with Python: A Guide for Data Scientists
Andreas C. Müller & Sarah Guido



Data Science from Scratch: First Principles with Python
Joel Grus

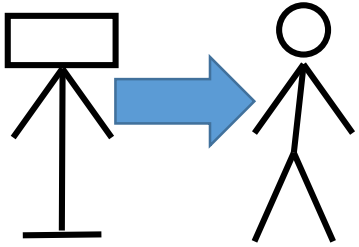


Ace the Data Science Interview: 201 Real Questions and Answers
KEVIN HUO & NICK SINGH



The Hundred-Page Machine Learning Book
Andriy Burkov

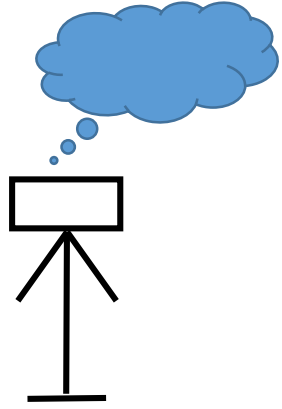
Your views



What would be an example of problematic manipulation?

- There is a tendency to recommend expensive books.
- There is a tendency to recommend books that other customers have bought.
- There is a tendency to recommend books that fit the interests of the user.
- There is a tendency to recommend books of a specific political persuasion.
- There is a tendency to recommend books from authors with diverse backgrounds.
- Only objective criteria are used to create recommendations, e.g. books that are about a similar topic.

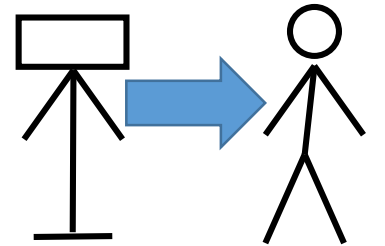
Philosophical input: a general idea



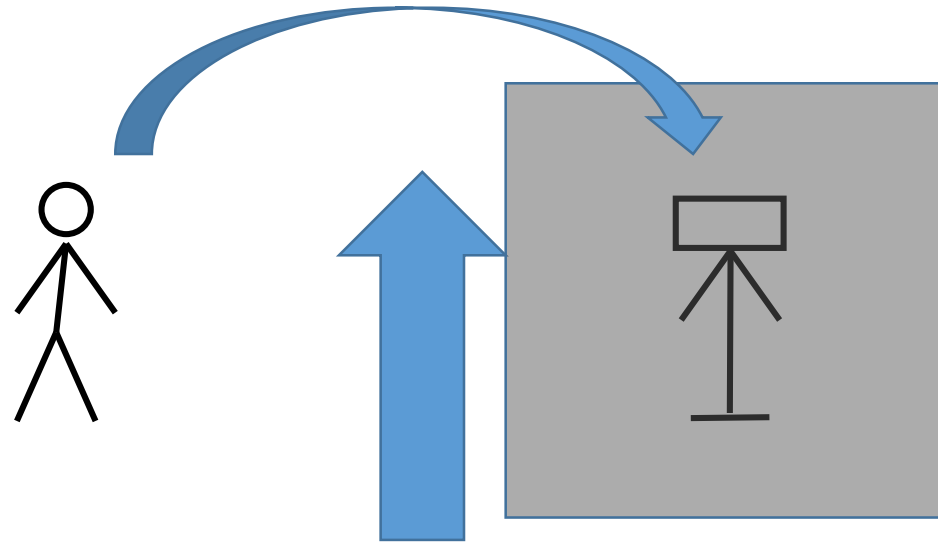
AI applications take decisions

No problem, if authorization by
„informed consent“

AI applications influence human decisions



Philosophical input: informed consent



Opacity

Rational basis: information

Philosophical input: opacity

OPENING THE BLACK BOX OF DEEP NEURAL NETWORKS VIA INFORMATION

Opening the black box of Deep Neural Networks via Information

Ravid Schwartz-Ziv

*Edmond and Lilly Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel*

RAVID.ZIV@MAIL.HUJI.AC.IL

Naftali Tishby*

*School of Engineering and Computer Science
and Edmond and Lilly Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel*

TISHBY@CS.HUJI.AC.IL

Editor: ICRI-CI

Abstract

Despite their great success, there is still no comprehensive theoretical understanding of learning with Deep Neural Networks (DNNs) or their inner organization. Previous work [Tishby and Zaslavsky (2015)] proposed to analyze DNNs in the *Information Plane*; i.e., the plane of the Mutual

“Despite their great success, there is still no comprehensive understanding of the optimization process or the internal organization of DNNs, and they are often criticized for being used as mysterious “black boxes””
p. 2

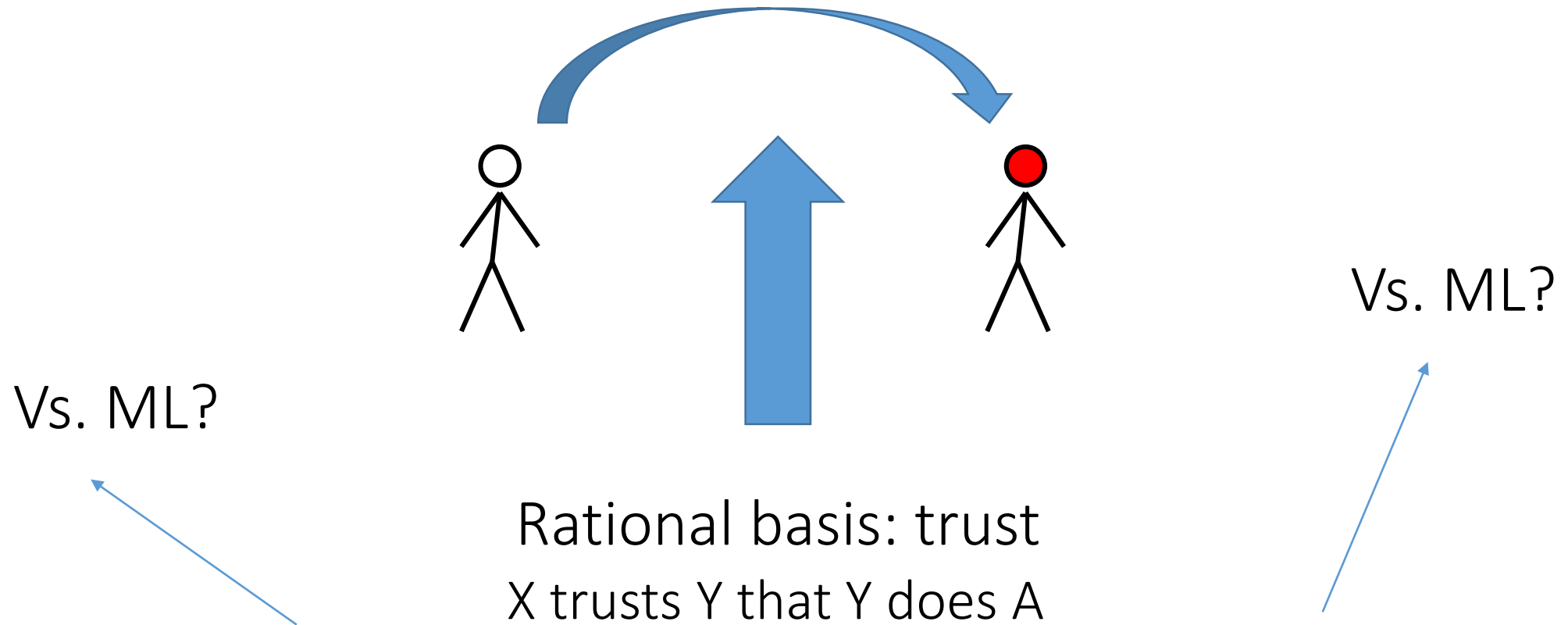
Shwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of Deep Neural Networks via Information.

[ArXiv:1703.00810](https://arxiv.org/abs/1703.00810).

Question for you

Is the opacity of ML models special? If so why?

Philosophical input: Trust



1. X thinks that Y does A.
2. X thinks that Y does so for goodwill.

Question for you

Under which conditions application
would **you** trust an ML application?

Philosophical input

Selected conditions:

- Human agency and oversight
- Transparency
- Accountability



Brussels, 8.4.2019
COM(2019) 168 final

**COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN
PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL
COMMITTEE AND THE COMMITTEE OF THE REGIONS**

Building Trust in Human-Centric Artificial Intelligence

[Communication: Building Trust in
Human Centric Artificial Intelligence](#)

Philosophical input: human agency and oversight

“Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Depending on the specific AI-based system and its application area, the appropriate degrees of **control measures**, including the adaptability, accuracy and explainability of AI-based systems, should be ensured¹². **Oversight** may be achieved through governance mechanisms such as ensuring a human-in-the-loop, human-on-the-loop, or human-in-command approach.¹³ It must be ensured that public authorities have the ability to exercise their oversight powers in line with their mandates. All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required.”

Philosophical input: transparency

“The **traceability** of AI systems should be ensured; it is important to log and document both the decisions made by the systems, as well as the entire process (including a description of data gathering and labelling, and a description of the algorithm used) that yielded the decisions. Linked to this, **explainability** of the algorithmic decision-making process, adapted to the persons involved, should be provided to the extent possible. Ongoing research to develop explainability mechanisms should be pursued.”

p. 5

Philosophical input: accountability

“**Potential negative impacts** of AI systems should be identified, assessed, documented and minimised. The use of impact assessments facilitates this process. These assessments should be proportionate to the extent of the risks that the AI systems pose. **Trade-offs** between the requirements – which are often unavoidable – should be addressed in a rational and methodological manner, and should be accounted for. Finally, when unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure **adequate redress**.”

p. 6

Question for you

What do you think about these requirements of

- Human agency and oversight
- Traceability
- accountability?

Summary

- Autonomy is a basic ethical concern.
 - If ML/AI's influence on humans is manipulative, autonomy is violated.
 - There are examples in which ML/AI manipulate human decisions.
 - Manipulation can be avoided by informed consent.
 - The basis for informed consent is trust.
 - Trust in ML is difficult due to its opacity.
 - It is debated when people should trust ML.
- Merci – thanks!