

Values in AI

many thanks to Aasa, Melanie and Sune for organizing!

fairmed.compute.dtu.dk

outline - the questions



shift of research focus to discuss *values* and *power*?

values, power are causes – while *fairness, bias* are effects / symptoms

are actions and values aligned ?

values cause future actions

massive misalignment in big tech actions and values

power analysis?

AI contributes to big tech power asymmetry

is an AI axiology research program lurking?

seems feasible to define and quantify AI's value system

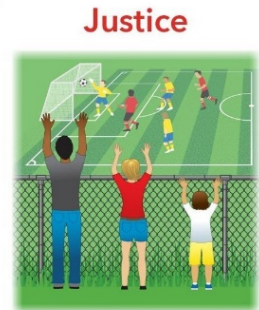
Utopian technical program: Safe AI!



The assumption is that **everyone benefits from the same supports**. This is equal treatment.



Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.



All 3 can see the game without supports or accommodations because **the cause(s) of the inequity was addressed**. The systemic barrier has been removed.

Meta-slide I:

Karen Hao interview w/ Joaquin Candela

Mar 11, 2021, in MIT Technology Review



“Joaquin Quiñonero Candela, a director of AI at Facebook, was apologizing to his audience. It was March 23, 2018, just days after the revelation that Cambridge Analytica...”

As he stepped up to face the room, he began with an admission. “I’ve just had the hardest five days in my tenure at Facebook,” he remembers saying. “If there’s criticism, I’ll accept it.”

[Facebook] algorithms were creating much faster, more personalized feedback loops for tweaking and tailoring each user’s news feed to keep nudging up engagement numbers....

Venturebeat: **“Hao described the subjects of her Facebook story as well-intentioned people trying to make changes in a rotten system that acts to protect itself.** Ethics researchers in a corporation of that size are effectively charged with considering society as a shareholder, but everyone else they work with is expected to think first and foremost about the bottom line, or personal bonuses. Hao said that reporting on the story has convinced her that self-regulation cannot work.”

<https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation>

<https://venturebeat.com/2021/03/12/ai-weekly-facebook-google-and-the-tension-between-profits-and-fairness/>

Meta-slide II:

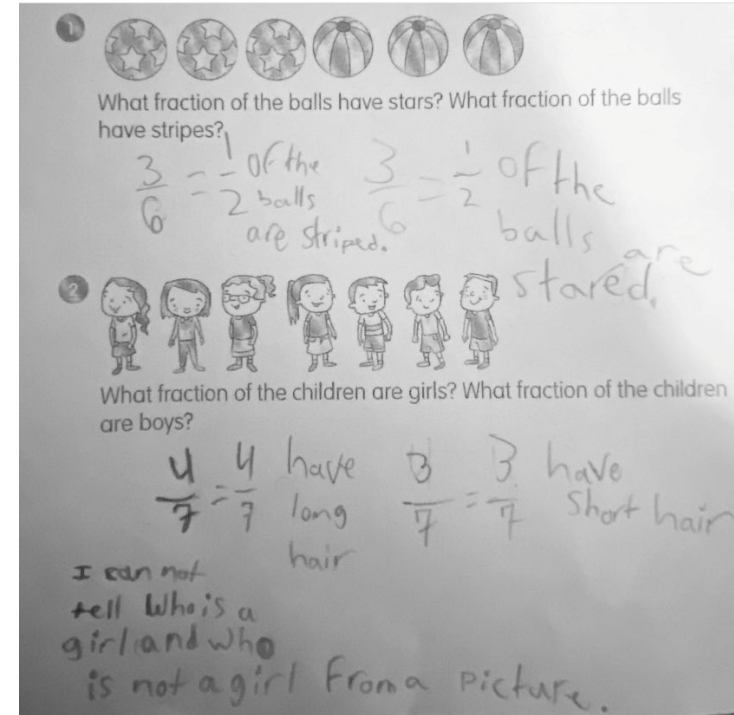
ML ethics nearly synonymous with *bias & privacy*

For good reasons!

Like most AI ethics issues
bias problems are not new

Tamar Rubinstein / @TamarPedsRheum
“My 8yo’s homework tonight” =>

But algorithms can lead to amplification
of bias problems ...examples are legio



Meta-slide III

UTOPISTS vs philosophers?

Xavier Denamur

LES PHILOSOPHES

Les citoyens en droit de savoir

Chers clients,

"C'est trop compliqué". Combien de fois j'ai pu entendre ces mots, depuis que je défends cette idée toute simple, que les



SALADES, TARTES CHAUDES ET PLATS VÉGÉTARIENS

Salade des Utopistes	€14.00
Panaché de crudités et Roquefort Carles assorti de chèvre chaud sur pain au levain	
Salade des Philosophes	€18.00
Panaché de crudités agrémenté de magret de canard fumé et chèvre chaud sur pain au levain et jambon Pata Negra Bellota	

Meta-slide IV Bits of personal history -AI ethics



Transparency and explainability in neurotechnology (1994 -)

- Uncertainty of explanations - NPAIRS with Steven Strother (2002-)
- Defense against explanation fairwashing w/ Laura Rieger: (2020)

Responsible business in the blogosphere (2009-2013)

- Measurement Systems for Ethical Capital in the Experience Economy. 12 mill DKK w/ Mette Morsing, CBS
- w/ Nicolai Peitersen's Actics – forward looking start-up - online rating tool for the ethics of corporations
- "Good friends, Bad News – Affect and virality in Twitter" w/ Adam Arvidsson, Finn Aarup
- Social media monitoring: "COP15 barometer" and a new tool for monitoring "Wikipedia edit wars"

Privacy for neuroinformatics (2015)

- Personal Data Storage for neurotech signals w/ Arek Stopczynski, Sune Lehmann, Sandy Pentland

Differential privacy (2018)

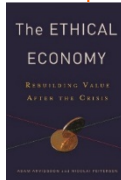
- for reporting student/task difficulty "Rasch model" w/ Teresa Steiner, David Nyrnberg

Safe AI proposal (2018)

- Part of our lobbying efforts to get the Danish Government to invest in AI

AI Pioner Center (2021)

- Serge Belongie's vision of Human Centered AI at scale: "Nothing about us without us"



Define AI: Intelligent systems... $\neq$ data fitting

Intelligent systems have senses

- Seek relevant data⁽¹⁾.. causal discovery, embodied
- Attention, multi-modal - not a one-trick pony

Intelligent systems are learning systems

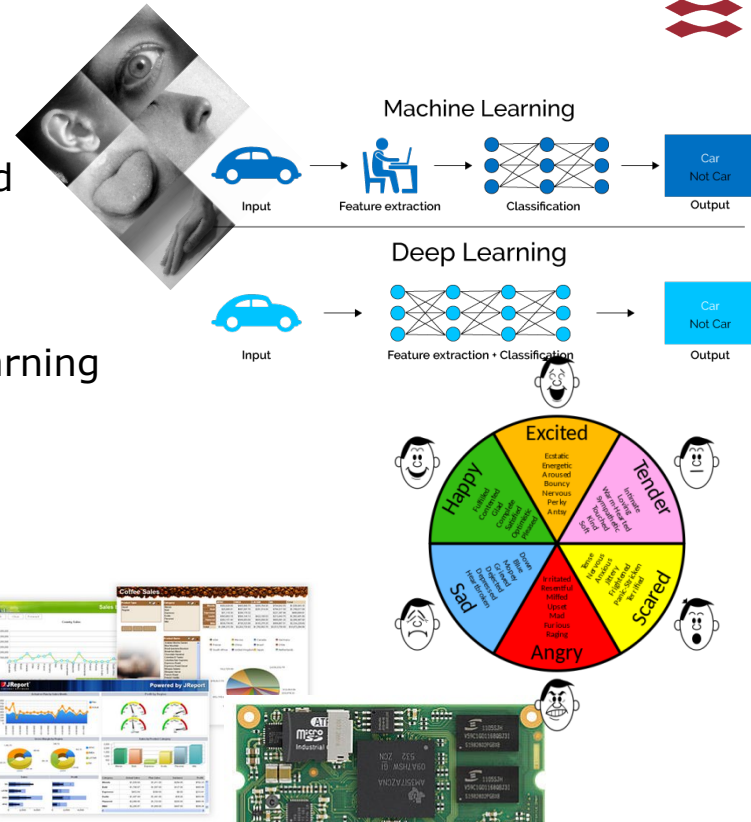
- Machine learning ... semi-supervised, transfer learning
- Active learning ... ask questions, intervention

Intelligent systems have social competences

- Communication, know own limitations
- Emotion detect, express, power

Intelligent systems perform "live"

- Global coordination – level C1 conscious ⁽²⁾
- Real-time operation -action perception ⁽³⁾



(1) Bajcsy, R., 1988. Active perception. Proceedings of the IEEE, 76(8), pp.966-1005.

(2) Dehaene, S., Lau, H. and Kouider, S., 2017. What is consciousness, and could machines have it?. *Science*, 358(6362), pp.486-492.

(3) Little, D.Y.J. and Sommer, F.T., 2013. Learning and exploration in action-perception loops. *Frontiers in neural circuits*, 7, p.37.

Define ethics, values



Ethics or moral philosophy is a branch of philosophy that "involves systematizing, defending, and recommending concepts of **right and wrong behavior**".

The field of ethics, along with aesthetics, concerning matters of value; these fields comprise the branch of philosophy called **axiology** (1)

In ethics, **value** denotes the degree of importance of some thing or action, with the aim of determining what actions are best to do or what way is best to live (normative ethics), or to describe the significance of different actions.

=> Values cause future actions

Philosophical value is distinguished from economic value, since it is independent from some other desired condition or commodity. (2)

Research & philosophical question:

Limits to **AI axiology** – is possible at all to quantify values?

Objects of **infinite** economic value? *Human life, Genomes, climate, cultural heritage*

Define power

=> Power predicts impact of actions



Oxford languages (1) **Noun**

1. the ability or capacity to do something or act in a particular way...
2. the capacity or ability to direct or influence the behaviour of others or the course of events...

CHAPTER 6

Value Creation and Power Asymmetries
in Digital Ecosystems: A Study of a Cloud
Gaming Provider

Arto Ojala, Nina Helander, and Pasi Tyrväinen

Astley and Sachdeva (1984) identified three sources of power:
hierarchical authority, **resource control**, and **network centrality**

Hierarchical authority often relates to official positions that actors have over one another, so they are usually coupled with actors like authorities or supervisors

Resource control looks at the environment of an organisation, as it states that everyone is dependent on the resources of others.

Network centrality, refers to the position of an actor in a network.

Asymmetrical power refers to a. relationship between two individuals in which one, the powerful person, has control over the outcomes of the other, the subordinate, but not vice versa (4)

(1) <https://languages.oup.com/google-dictionary-en/>

(2) Ojala, A., Helander, N. and Tyrväinen, P., 2020. Value Creation and Power Asymmetries in Digital Ecosystems: A Study of a Cloud Gaming Provider. In *Measuring the Business Value of Cloud Computing* (pp. 89-106). Palgrave Macmillan, Cham.

(3) Astley, W.G. and Sachdeva, P.S., 1984. Structural sources of intraorganizational: Power: A theoretical synthesis. *Academy of management review*, 9(1), pp.104-113.

(4) Goodwin, S.A., 1993. Impression formation in asymmetrical power relationships: does power corrupt absolutely?

First order AI power problem

**Agent = AI system
can control your actions**

**Agent's values are not aligned
with yours...**

...uncool job, slavery,

A personal take on science and society

World view

Don't ask if AI is good or fair, ask how it shifts power



By Pratyusha
Kalluri

Pratyusha Kalluri is a co-creator of the Radical AI Network and an AI researcher at Stanford University in California.
e-mail: pkalluri@stanford.edu

Through the lens of power, it's possible to see why accurate, generalizable and efficient AI systems are not good for everyone. In the hands of exploitative companies or oppressive law enforcement, a more accurate facial recognition system is harmful. Organizations have responded with pledges to design 'fair' and 'transparent' systems, but fair and transparent according to whom?

These systems sometimes mitigate harm, but are controlled by powerful institutions with their own agendas. At best, they are unreliable; at worst, they masquerade as 'ethics-washing' technologies that still perpetuate inequity.

Power analysis:

Are Facebook behaviors aligned with their values?

Are Facebook behaviors aligned with your values?

Is the FB-you relation power symmetric – terms negotiated?

Kasy and Abebe: Fairness, equality, and power in algorithmic decision-making.



Much of the debate on the impact of algorithms is concerned with fairness, defined as the absence of discrimination for individuals with the same "merit."

Drawing on the theory of justice, we argue that leading notions of fairness are limited:

- Legitimize inequalities justified by merit;
- Narrowly bracketed, considering only differences of treatment within the algorithm; and they consider between-group and not within group differences ("identity politics")
- Contrast fairness-based perspective with two alternate perspectives:
 - Focus on inequality and the causal impact of algorithms on the distribution of power.

Kasy and Abebe use the insights to present a guide for algorithmic auditing & discuss the importance of inequality and power-centered frameworks in algorithmic decision-making.

Kasy, M. and Abebe, R., 2021, March. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 576-586).

Ethics of AI: Benefits and risks of artificial intelligence

Tiernan Ray ZDNet Apr 30 2021



"Ethics in AI is essentially questioning, constantly investigating, and never taking for granted the technologies that are being rapidly imposed upon human life.

That questioning is made all the more urgent because of *scale*. AI systems are reaching tremendous size in terms of the compute power they require, and the data they consume. And their prevalence in society, both in the scale of their deployment and the level of responsibility they assume, dwarfs the presence of computing in the PC and Internet eras.

At the same time, increasing scale means many aspects of the technology, especially in its deep learning form, escape the comprehension of even the most experienced practitioners.

Ethical concerns range from the esoteric, such as who is the author of an AI-created work of art; to the very real and very disturbing matter of surveillance in the hands of military authorities who can use the tools with impunity to capture and kill their fellow citizens."

AI with quantifiable values



Explicit value systems?

Sustainable Development Goals (SDGs)

AI4Good: "What if AI were developed to serve humanity rather than commerce?"

Implicit value statements?

- Reconstruct values from actions



AI for Good

Google search

Sales, speed, findability vs Unbiased retrieval of relevant docs, availability

Facebook:

Time spent at FB vs social interaction and truths

Amazon, Nemlig.com

Profit vs fair deals, workplace ethics

The Guardian, Sep 19, 2017



Power analysis – value-action alignment I



Google values: "Ten things we know to be true..."

- Focus on the user and all else will follow. ...
- It's best to do one thing really, really well. ...
- Fast is better than slow. ...
- **Democracy** on the web works. ...
- You don't need to be at your desk to need an answer. ...
- You can make money **without doing evil**. ...
- There's always more information out there

On diversity:

Google is committed to continuing to make diversity, equity, and inclusion part of everything we do—from how we build our products to how we build our workforce.

Google is growing to fulfill that vision <https://diversity.google>

Power analysis – value-action alignment II



Ads by Google

[Latonya Evans, Arrested?](#)

1) Enter Name and State. 2) Access Full Background Checks Instantly.

www.instantcheckmate.com/

Ads by Google

[Latisha Smith Located](#)

Background Check, Arrest Records, Phone, & Address. Instant, Accurate

www.instantcheckmate.com/

Adverse messaging for black-identifying names.

Ads by Google

[Latonya Evans's Records](#)

1) Enter Name and State. 2) Access Full Background Checks Instantly.

www.instantcheckmate.com/

Ads by Google

[Latisha smith: Truth](#)

Arrests and Much More. Everything About Latisha smith

www.instantcheckmate.com/

Examples of ads for "Latonya Evans," "Latisha Smith" — (Sweeney, 2013)

Remind:

AI is "live" – continued learning from users – auditing is a continual process

AI is not classical data fitting – no fixed & curated training+test sets



Rachel Thomas @math_Rachel

Co-founder <http://fast.ai> | past: founding director of USF Center for Applied Data Ethics



Warehouse automation has transformed warehouse workers into adjuncts for robots, not the other way around.

The robots set the pace, literally: "The average worker picks roughly 100 items per hour if walking around, but more than 300 items an hour in the automated system."

This pace is set by the robot, and the repetitive, high intensity standing, bending and reaching labor has caused injury rates to increase every year in proportion to the degree of automation in Amazon warehouses:

Misalignment of big tech actions vs values!

Berlingske

NYHEDER

OPINION

BUSINESS

ARK

DANMARK

Nemlig.com straffer chauffører med bøder for små forsinkelser

Dagens overblik: Berlingske giver dig her overblik over dagens vigtigste historier - om den vedholdende coronakrise og de øvrige store begivenheder og samtaleemner i ind- og udland.



Tvang og fysisk afstraffelse: Fysisk afstraffelse, trusler, vold, eller anden form for psykisk eller fysisk tvang eller misbrug accepteres ikke af Nemlig.com

<https://www.nemlig.com/om-nemlig/baeredygtighed/ansvarlighed/code-of-conduct>

Second order AI power problem

**Agent = AI system
can control your actions &
agent's explicit values are aligned
with yours...**

Yet,

**Agent's control actions are
misaligned with explicit values
(implicit values $\neq$ explicit values)**

Utopian fixes?

Resistance AI Workshop – NeurISP 2020

“The goal of the Resistance AI Workshop is to examine how AI shifts power and how we can build human/AI systems that shift power to the people.”

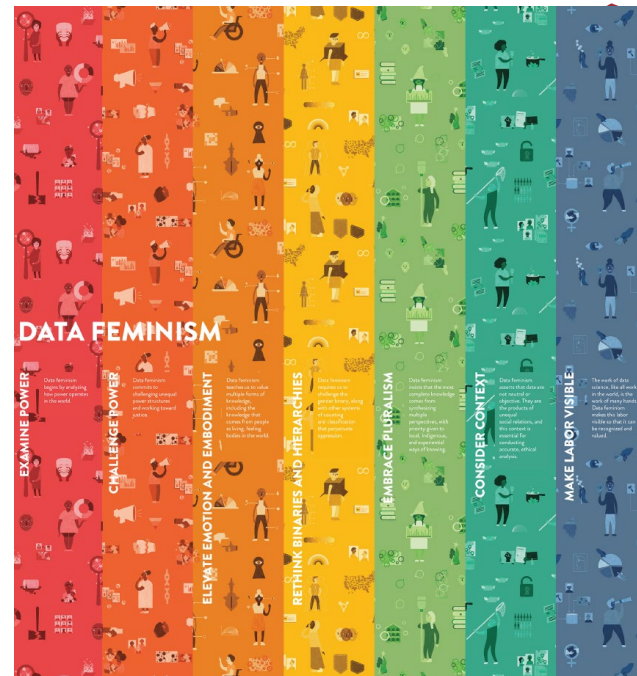
Excerpts from program:

Catherine D'Ignazio, Lauren F Klein, Marcia Diaz,
“The Data Feminism Infographic”

Stefano Diana,

“Rewriting Marx to understand AI and the data society.”

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew,
Ravit Dotan, Michelle Bao,
“Values of Machine Learning”



<https://sites.google.com/view/resistance-ai-neurips-20>

AI Ethics Guidelines Global Inventory



<https://inventory.algorithmwatch.org/>

United Nation SDGs = universal representation of values?

PERSPECTIVE

<https://doi.org/10.1038/s41467-019-14108-y>

OPEN

The role of artificial intelligence in achieving the Sustainable Development Goals

Ricardo Vinuesa^{1*}, Hossein Azizpour², Iolanda Leite², Madeline Balaam³, Virginia Dignum⁴, Sami Domisch⁵, Anna Felländer⁶, Simone Daniela Langhans^{7,8}, Max Tegmark⁹ & Francesco Fuso Nerini^{10*}

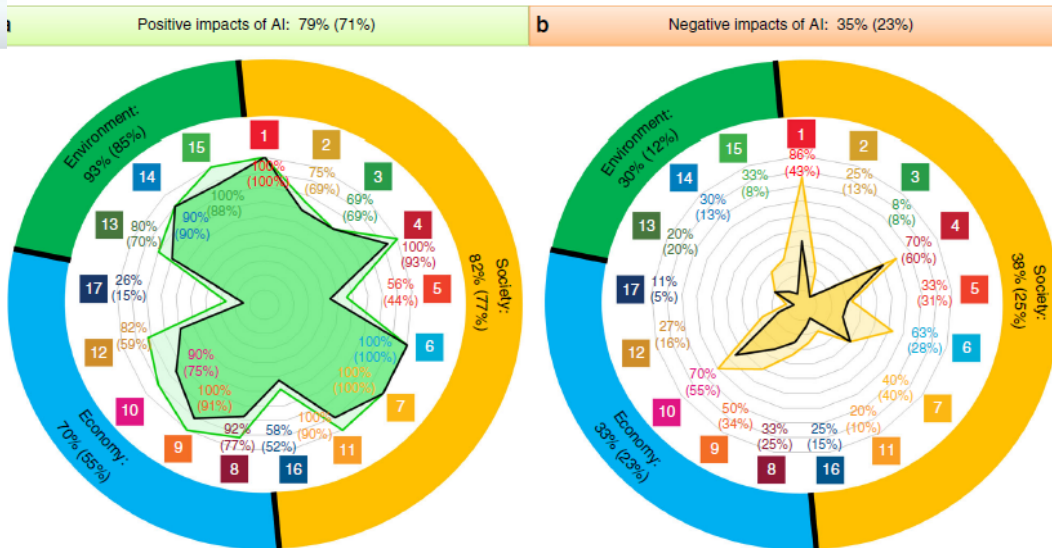
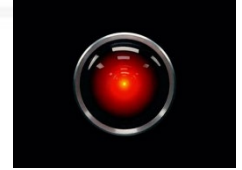
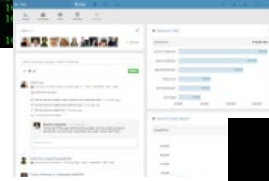
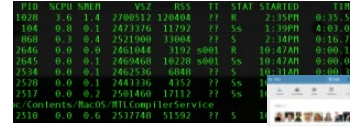


Fig. 1 Summary of positive and negative impact of AI on the various SDGs. Documented evidence of the potential of AI acting as (a) an enabler or (b) an inhibitor on each of the SDGs. The numbers inside the colored squares represent each of the SDGs (see the Supplementary Data 1). The percentages on the top indicate the proportion of all targets potentially affected by AI and the ones in the inner circle of the figure correspond to proportions within each SDG. The results corresponding to the three main groups, namely Society, Economy, and Environment, are also shown in the outer circle of the figure. The results obtained when the type of evidence is taken into account are shown by the inner shaded area and the values in brackets.

Safe AI - checklist



Safe AI = secure – test & verified software and hardware

Safe AI = open source – methods, code, hardware

Safe AI = self-conscious – understands own role in creating reality

Safe AI = can keep a secret – privacy by design

Safe AI = has calibrated values – SDGs, debug for stereotypes, bias

Safe AI = is accountable - transparent, communicating - explains

Safe AI = understands social relations – navigate by user's knowledge graph

Safe AI = understands power - negotiate from a symmetric position

Safe AI = generates trust



Take home questions

shift of research focus to discuss *power*?

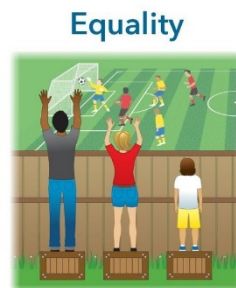
as power is a cause – while *fairness*, *bias* are effects / symptoms
are actions and values aligned ?

massive misalignment in big tech actions and values
power analysis?

big tech AI contributes massively to power asymmetry
is an AI axiology research program lurking?

seems feasible to define and infer AI values
from actions – alignment metrics?

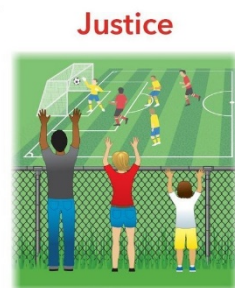
Utopian technical program: Safe AI?



The assumption is that
**everyone benefits from
the same supports.** This
is equal treatment.



**Everyone gets the
supports they need**
(this is the concept of
"affirmative action"), thus
producing equity.



All 3 can see the game
without supports or
accommodations because
**the cause(s) of the
inequity was addressed.**
The systemic barrier has
been removed.