hcaim

human centred
artificial intelligence
masters

ETHICAL ASPECTS IN RESEARCH TOPICS

Mihály Héder

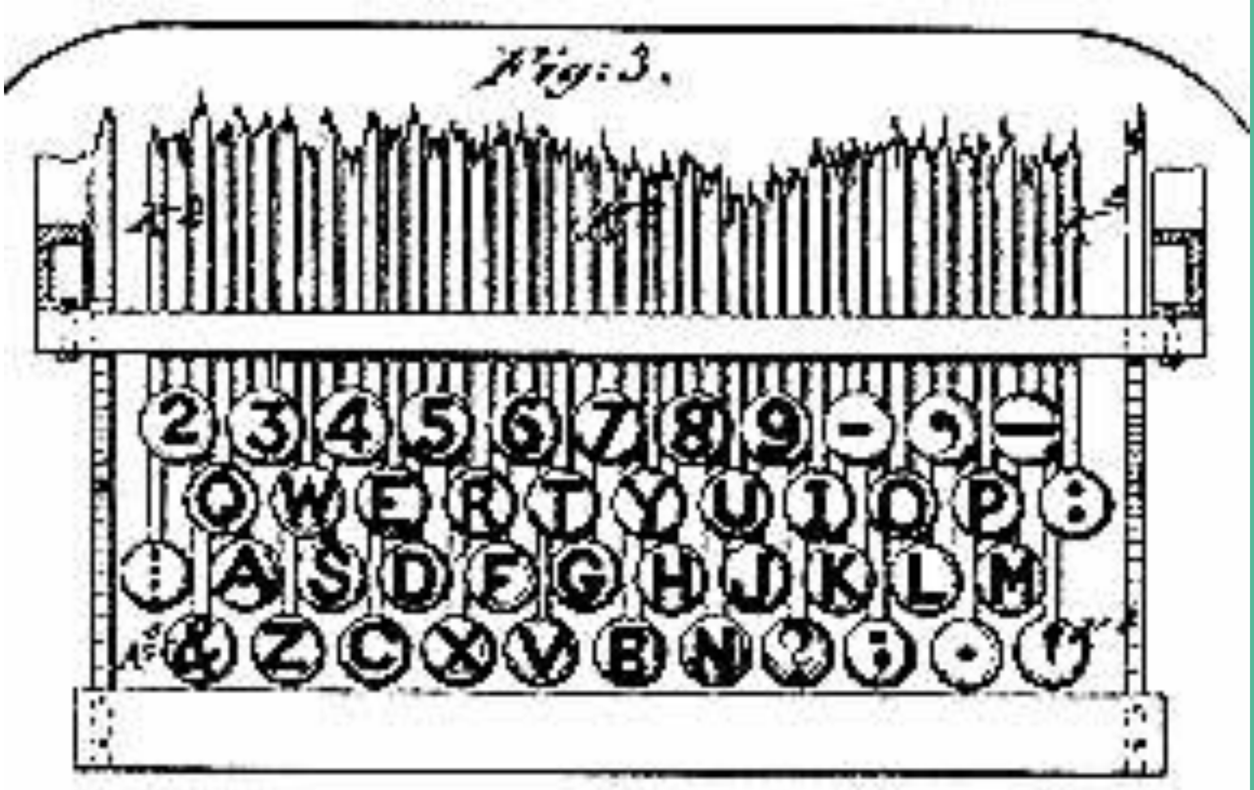
Department of Philosophy and History of Science
www.filozofia.bme.hu
Budapest University of Technology and Economics

HU Utrecht, 27 January 2025.

humancentered-ai.eu

WHY are we here? (on a human-centered AI event)?

The history of the QWERTY keyboard layout



The first typewriter

- The first typewriter was created by Christopher Latham Sholes in the 1860s
- There were a lot of technological problems:
    - The typewriter easily jammed, especially if the user was fast enough
- Sholes started with the alphabetic order first
    - He re-arranged the keyboard to a random layout (QWERTY) in order to ***slow down the user***

QWERTY's diffusion

- In the **1880s** there were concurrent keyboard layouts, QWERTY was one of them
- By 1890 there was no technological reason to use QWERTY
  - But it was bought by Remington and spread everywhere called "universal"

Remington

STANDARD

Typewriter.

WYCKOFF, SEAMANS & BENEDICT, 327 Broadway, New York, Boston, Mass.; Philadelphia, Pa.; Washington, D. C., Baltimore, Md.; Chicago, Ill.; St. Louis, Mo.; Indianapolis, Ind.; Minneapolis, Minn.; St. Paul, Minn.; Kansas City, Mo.; Denver, Col.; Cincinnati and Cleveland, O.; London, England.

NOTE. — *Our unqualified challenge for a test of all Writing Machines remains unaccepted.* Send for copy if interested.

[1888]

Locked in to QWERTY

- By the 20th century society was locked in to QWERTY
- There are three reasons of this
  - Economics of scale (mass manufacturing)
  - Interdependence of technology
  - Irredeemable investments

Collingridge-dilemma

- There lies a deep tension in the logic of technology development:
    - When a **technology is new** it is easy to modify
        - In theory, big problems could be avoided
        - BUT there is not enough information to discover the exact problems
    - In case of an **established, ubiquitous technology** it is easy to see the problems
        - But it is **hard to make changes**

example: Dichlorodiphenyltrichloroethane (DDT)

AI vs. technological lock-in

Why do we need to deal with AI specifically?

- **Potential technology lock-in:**
  - The impact is greater: fewer and fewer **technology owners** can influence a larger and larger slice of the world
  - AI is **software**
    - consequently: the marginal cost of its "multiplication" is negligible.
    - e.g. if there will be an effective AI diagnostic system / psychologist / accountant it will be more worthwhile to copy it than to rewrite it

Why do we need to deal with AI specifically?

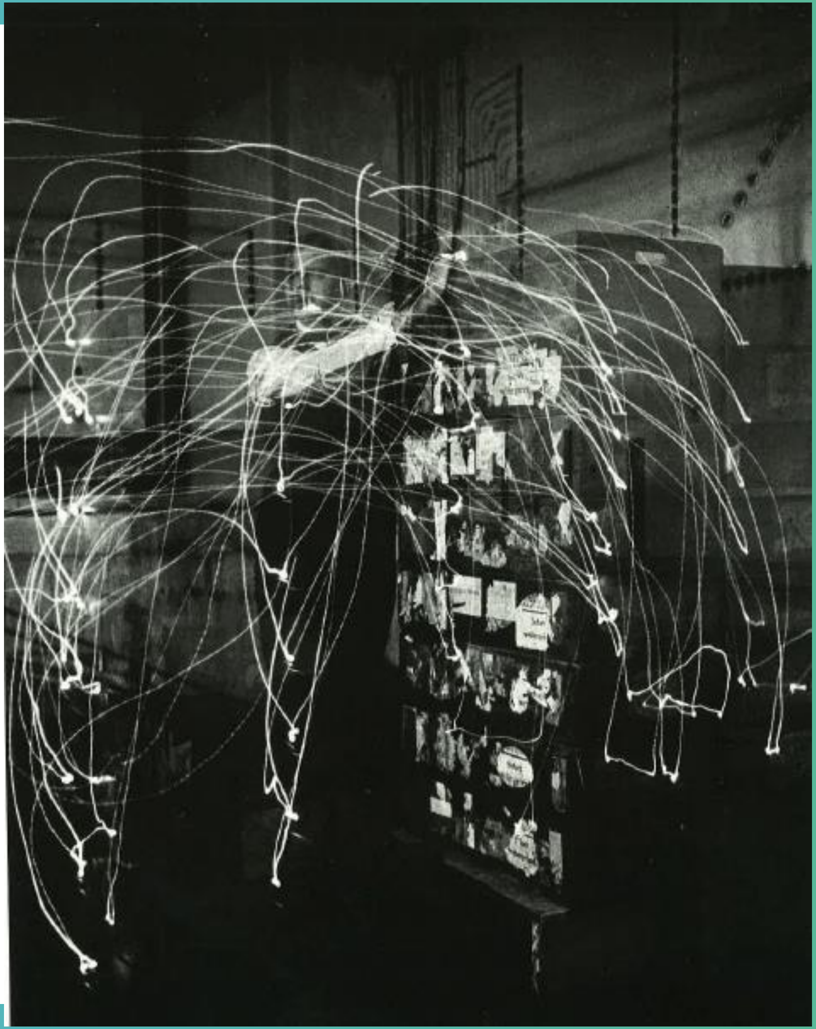
- **Potential technology lock-in:**
  - Moreover, AI is Software-as-a-Service, so it can be provided from anywhere in the world in most cases
    - except locally controlled robots
  - AI developers start in a certain direction, which will be difficult to change later
    - similar situation: e.g. gmail accounts - raise your hand if you don't have one

Solution attempt #1
Human-Centered Design of machines?

What is not human-centered design?

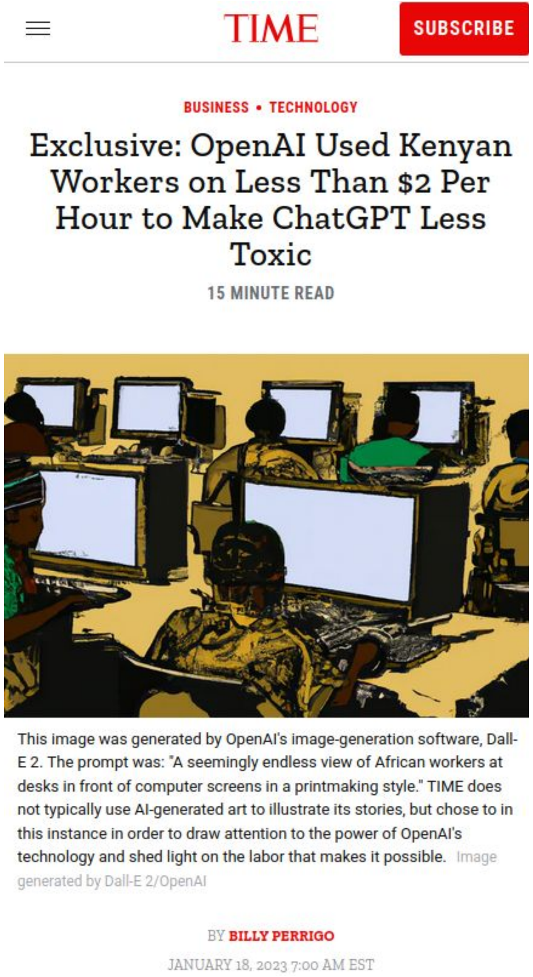
Source: Elias Beck. *'Child Labor in the Industrial Revolution'*. History Crunch. December 30, 2021. https://www.historycrunch.com/child-labor-in-the-industrial-revolution.html#/

What is not human-centered design?



Gilbreth chronocyclograph of motions necessary to move and file sixteen boxes full of glass, n.d. From: Mike Mandel, *Making Good Time: Scientific Management, the Gilbreths, Photography and Motion, Futurism* (Santa Cruz, CA: California Museum of Photography, University of California, Riverside, 1989), 26.

What is not human-centered design?



TIME

SUBSCRIBE

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

15 MINUTE READ

This image was generated by OpenAI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible. Image generated by Dall-E 2/OpenAI

BY BILLY PERRIGO

JANUARY 18, 2023 7:00 AM EST

hcaim human centred artificial intelligence masters

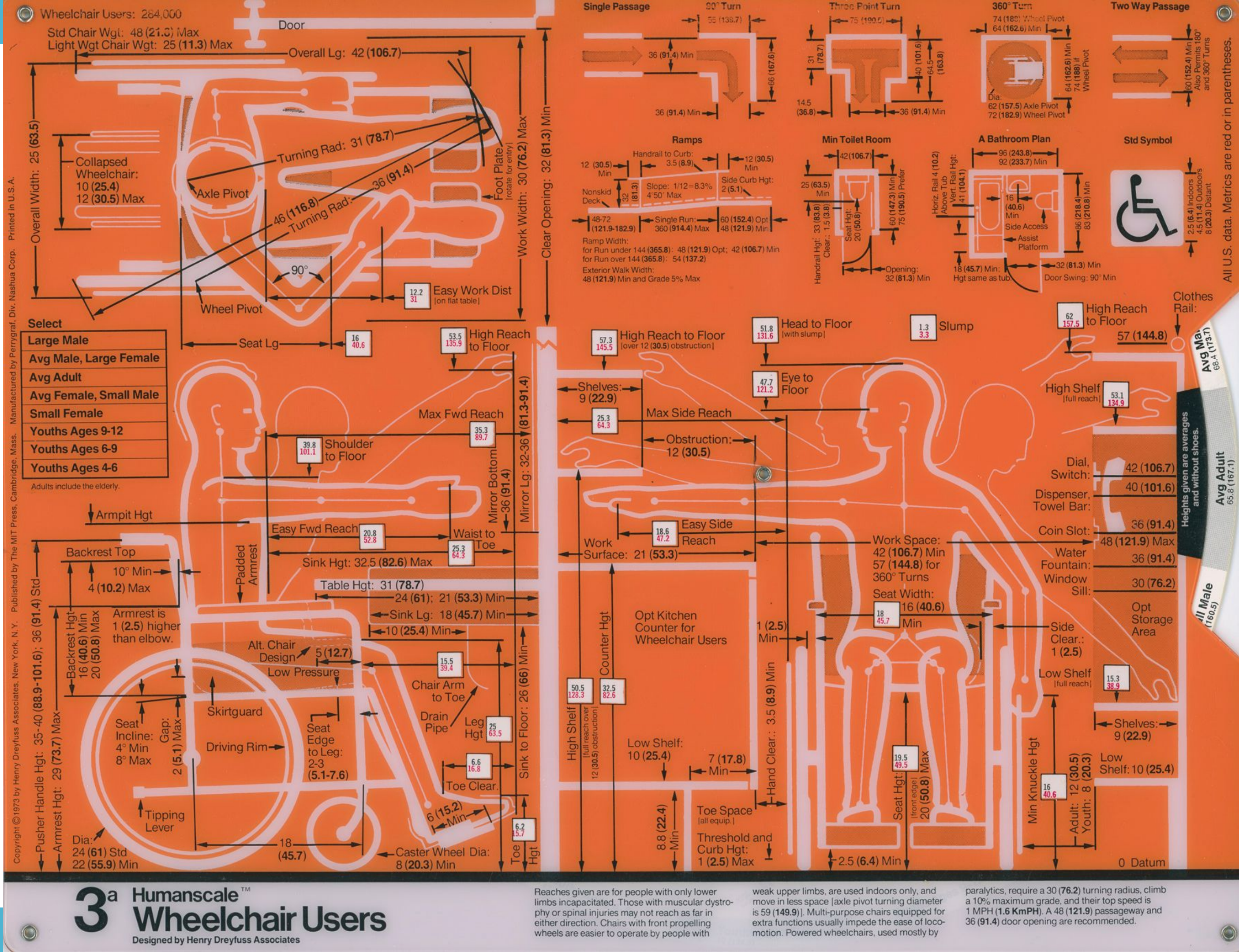
What is the machine maker's role on all of this?

- Self-absolving strategy 1
  - "The imperative of technology"
    - This is what **efficiency dictates**
- Self-absolving strategy 2
  - This is **what the customer wants**
- Self-absolving strategy 3
  - It **was legal** at the time
- Human-Centered Design: **rejection of all the above**
  - **humanities toolkit**
    - **argumentation**
    - **critical thinking**

Knowledge about Humans

Image: Henry Dreyfuss Associates, *Humanscale* selector 3a “Wheelchair Users,” 1974. Plastic, paper, and metal. Milwaukee Art Museum Research Center.

Source: Hanna Pivo, 20th-Century Tools for Measuring Time and Bodies April 19, 2019 https://blog.mam.org/2019/04/19/20th-century-tools-for-measuring-time-and-bodies/



Wheelchair Users: 284,000
Std Chair Wgt: 48 (21.8) Max
Light Wgt Chair Wgt: 25 (11.3) Max
Door
Overall Lg: 42 (106.7)
Overall Width: 25 (63.5)
Collapsed Wheelchair: 10 (25.4) 12 (30.5) Max
Axle Pivot
Turning Rad: 31 (78.7)
36 (91.4)
46 (116.8) Turning Rad:
90°
Wheel Pivot
Foot Plate [rotate for entry]
Work Width: 30 (76.2) Max
Clear Opening: 32 (81.3) Min
12.2 31 Easy Work Dist [on flat table]
Seat Lg
16 40.6
53.5 135.9 High Reach to Floor
Select
Large Male
Avg Male, Large Female
Avg Adult
Avg Female, Small Male
Small Female
Youths Ages 9-12
Youths Ages 6-9
Youths Ages 4-6
Adults include the elderly.
39.8 101.1 Shoulder to Floor
Max Fwd Reach
35.3 89.7
Mirror Bottom: 36 (91.4)
Mirror Lg: 32-36 (81.3-91.4)
Armpit Hgt
Easy Fwd Reach 20.8 52.8
Waist to Toe 25.3 64.3
Backrest Top
10° Min
4 (10.2) Max
Padded Armrest
Sink Hgt: 32.5 (82.6) Max
Table Hgt: 31 (78.7)
24 (61); 21 (53.3) Min
Sink Lg: 18 (45.7) Min
10 (25.4) Min
Armrest is 1 (2.5) higher than elbow.
Backrest Hgt: 16 (40.6) Min 20 (50.8) Max
Pusher Handle Hgt: 35-40 (88.9-101.6); 36 (91.4) Std
Armrest Hgt: 29 (73.7) Max
Alt. Chair Design 5 (12.7) Low Pressure
15.5 39.4 Chair Arm to Toe
Skirtguard
Seat Incline: 4° Min 8° Max
Gap: 2 (5.1) Max
Driving Rim
Seat Edge to Leg: 2-3 (5.1-7.6)
Drain Pipe
Leg Hgt 25 63.5
6.6 16.8 Toe Clear.
Sink to Floor: 26 (66) Min
Tipping Lever
Dia: 24 (61) Std 22 (55.9) Min
18 (45.7)
6 (15.2) Min
Caster Wheel Dia: 8 (20.3) Min
6.2 15.7 Toe Hgt
Copyright ©1973 by Henry Dreyfuss Associates. New York, N.Y. Published by The MIT Press, Cambridge, Mass. Manufactured by Perrygraf, Div. Nashua Corp. Printed in U.S.A.
Single Passage
90° Turn
36 (91.4) Min
Three Point Turn
360° Turn
74 (188) Wheel Pivot
64 (162.6) Min
Dia: 62 (157.5) Axle Pivot 72 (182.9) Wheel Pivot
Two Way Passage
60 (152.4) Min Also Permits 180° and 360° Turns
Ramps
Handrail to Curb: 3.5 (8.9)
12 (30.5) Min
Nonskid Deck
Slope: 1/12=8.3% 4'50' Max
Side Curb Hgt: 2 (5.1)
Single Run: 360 (914.4) Max
60 (152.4) Opt 48 (121.9) Min
48-72 (121.9-182.9)
Ramp Width: for Run under 144 (365.8): 48 (121.9) Opt; 42 (106.7) Min for Run over 144 (365.8): 54 (137.2)
Exterior Walk Width: 48 (121.9) Min and Grade 5% Max
Min Toilet Room
42 (106.7)
25 (63.5) Min
Handrail Hgt: 33 (83.8)
Seat Hgt: 20 (50.8)
60 (147.3) Min 75 (190.5) Prefer
Opening: 32 (81.3) Min
A Bathroom Plan
96 (243.8) 92 (233.7) Min
Horiz. Rail 4 (10.2) Above Tub
16 (40.6) Min Side Access
Assist Platform
18 (45.7) Min; Hgt same as tub
32 (81.3) Min Door Swing: 90° Min
Std Symbol
2.5 (6.4) Indoors 4.5 (11.4) Outdoors 8 (20.3) Distant
All U.S. data Metrics are red or in parentheses.
57.3 145.5 High Reach to Floor [over 12 (30.5) obstruction]
51.8 131.6 Head to Floor [with slump]
1.3 3.3 Slump
62 157.5 High Reach to Floor
Clothes Rail:
57 (144.8)
Shelves: 9 (22.9)
25.3 64.3 Max Side Reach
47.7 121.2 Eye to Floor
Obstruction: 12 (30.5)
High Shelf [full reach] 53.1 134.9
Dial, Switch: 42 (106.7)
Dispenser, Towel Bar: 40 (101.6)
Coin Slot: 36 (91.4)
48 (121.9) Max
Water Fountain: 36 (91.4)
Window Sill: 30 (76.2)
18.6 47.2 Easy Side Reach
Work Surface: 21 (53.3)
Work Space: 42 (106.7) Min 57 (144.8) for 360° Turns
Seat Width: 16 (40.6) 18 45.7 Min
Opt Kitchen Counter for Wheelchair Users
Counter Hgt 32.5 82.6
High Shelf [full reach over 12 (30.5) obstruction] 50.5 128.3
1 (2.5) Min
Side Clear.: 1 (2.5)
Opt Storage Area
Low Shelf [full reach] 15.3 38.9
Low Shelf: 10 (25.4)
7 (17.8) Min
Hand Clear.: 3.5 (8.9) Min
19.5 49.5 Seat Hgt [front edge] 20 (50.8) Max
Shelves: 9 (22.9)
Min Knuckle Hgt 16 40.6
Low Shelf: 10 (25.4)
Adult: 12 (30.5) Youth: 8 (20.3)
8.8 (22.4) Min
Toe Space [all equip.]
Threshold and Curb Hgt: 1 (2.5) Max
2.5 (6.4) Min
0 Datum
Avg Male 68.4 (173.7)
Avg Adult 65.8 (167.1)
Small Male (160.5)
Heights given are averages and without shoes.
3a Humanscale™ Wheelchair Users
Designed by Henry Dreyfuss Associates
Reaches given are for people with only lower limbs incapacitated. Those with muscular dystrophy or spinal injuries may not reach as far in either direction. Chairs with front propelling wheels are easier to operate by people with weak upper limbs, are used indoors only, and move in less space [axle pivot turning diameter is 59 (149.9)]. Multi-purpose chairs equipped for extra functions usually impede the ease of locomotion. Powered wheelchairs, used mostly by paralytics, require a 30 (76.2) turning radius, climb a 10% maximum grade, and their top speed is 1 MPH (1.6 KmPH). A 48 (121.9) passageway and 36 (91.4) door opening are recommended.

Ethical topics in AI

AI Ethics Problem Set:

THE FUTURE OF WORK

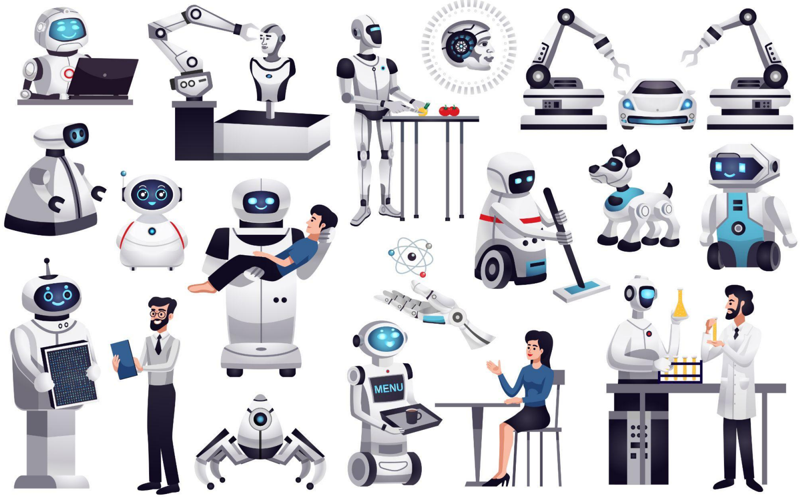


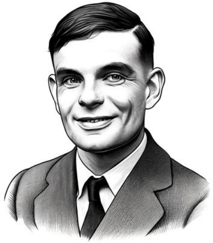
Image credit: Freepik premium

Why is AI special?
Future of Work

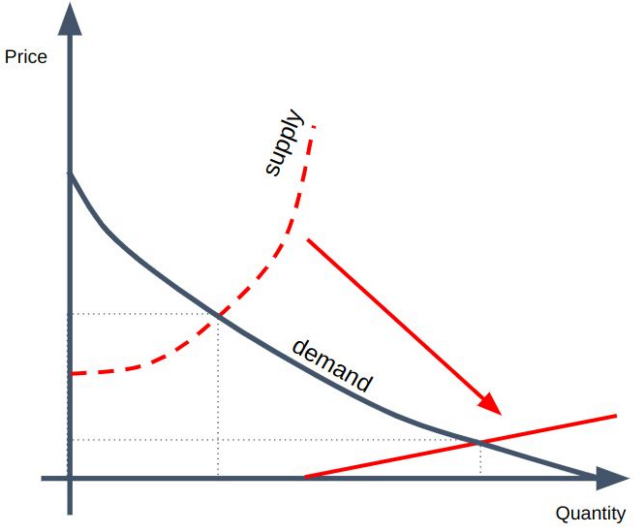
- AI threatens jobs
- This could result in another economic transformation similar to the **industrial revolutions**
  - from the point of view of today's people, the industrial revolutions represent the historical antecedent of a comfortable, technological life,
  - but those who experienced it could also be accompanied by serious disruption and impoverishment

Problem Set: The Artificial Worker

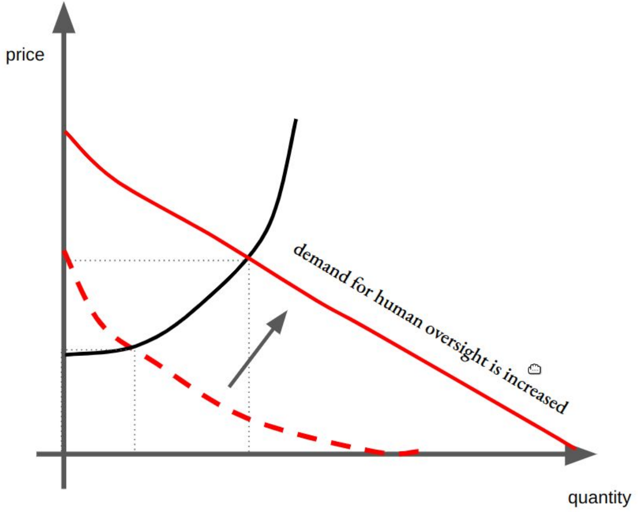
human centred
artificial intelligence
masters



new market equilibrium of junior work



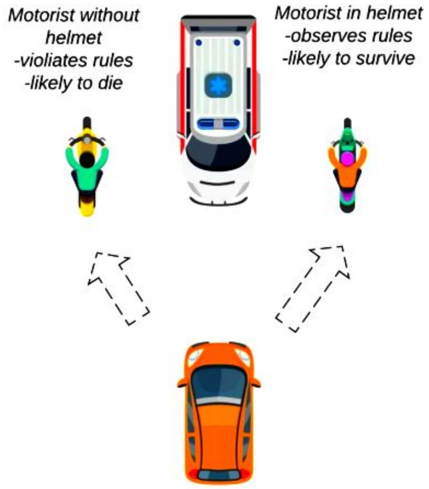
new market eq for oversight



Machine Ethics

AI Ethics Problem Set:

MACHINE ETHICS



P. Lin helmet problem

Image: M Héder & Dreamstime stock illustrations
Thought experiment by P. Lin. See:
Héder, M. The epistemic opacity of autonomous systems and the ethical consequences. *AI & Soc* (2020).
https://doi.org/10.1007/s00146-020-01024-9 (CC 4.0)

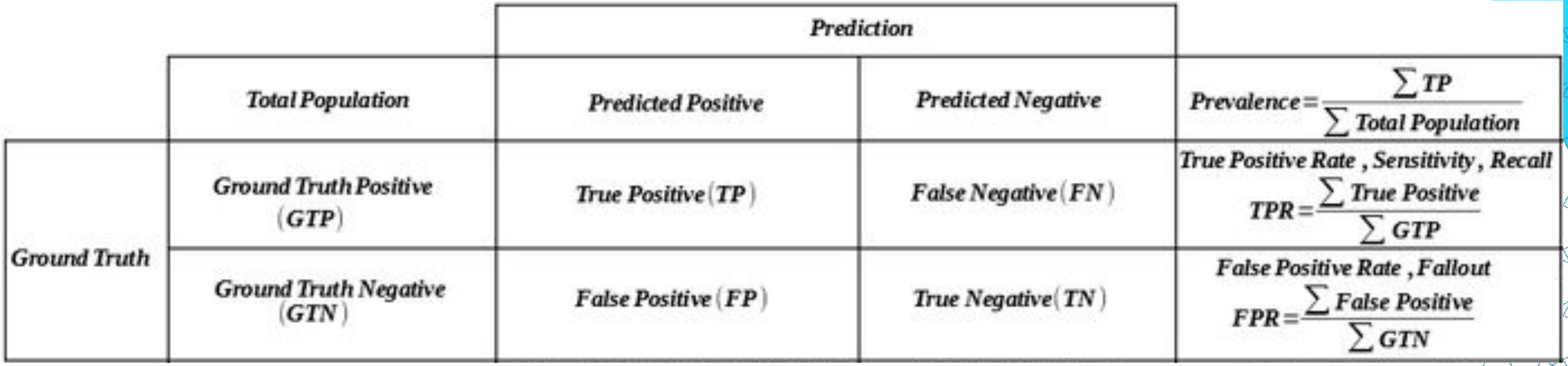
Why is AI special? Machine Ethics

- Designing the behavior of an artificial person (machine ethics)
  - we give the machine autonomy, since we precisely want it to think and make decisions for us
    - analogy: *raising children*
    - we want to avoid: biases, opacity
  - We place a machine in a "trusted" position in an unprecedented way
  - It arises that that artificial agent develops its own goals and does not only serve our goals

Topics of machine ethics

- Transparency
  - explainability
- Fairness
  - bias
  - lack of discrimination
- Alignment
  - wider social **values** like wellbeing
    - brings in politics

Fairness metrics



		Prediction		
	Total Population	Predicted Positive	Predicted Negative	$Prevalence = \frac{\sum TP}{\sum Total\ Population}$
Ground Truth	Ground Truth Positive (GTP)	True Positive (TP)	False Negative (FN)	True Positive Rate, Sensitivity, Recall $TPR = \frac{\sum True\ Positive}{\sum GTP}$
	Ground Truth Negative (GTN)	False Positive (FP)	True Negative (TN)	False Positive Rate, Fallout $FPR = \frac{\sum False\ Positive}{\sum GTN}$

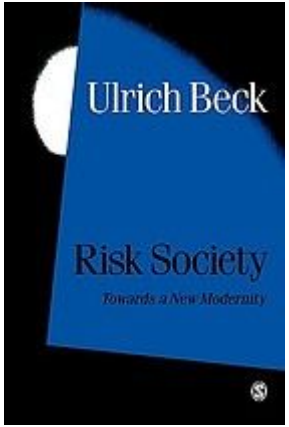
- Equal Opportunity
  - TPR_0 ~= TPR_1
- Equalized Odds
  - TPR_0 ~= TPR_1
  - FPR_0 ~= FPR_1

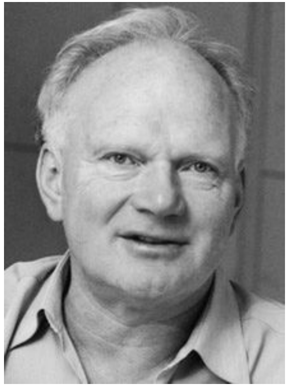
Demand from society

The social control of technology?

Ulrich Beck’s analysis (1980’s)

- A main feature of modern society is that it is **preoccupied with the future**, and
    - especially the negative scenarios, that is ‘risks’
- Catastrophes were formerly attributed to bad luck or divine acts
    - but not in Humanity’s control
- Now that our control seems greater (modern science) the **responsibility is ours**
    - this in turn undermines the institutions of modern society, e.g. trust in science





1944-2015

Modernity 2 (Beck's society)

- The victories of the first modernity (taking risks) have a boomerang effect
- Taking risks not serve us anymore
- So we enter **reflexive modernity**
    - Pesticide
    - Ozone
    - Nuclear
    - Toxins
    - CFC
    - Plastics, etc.
- Plus, we anticipate even more negative consequences
    - AI, GMO

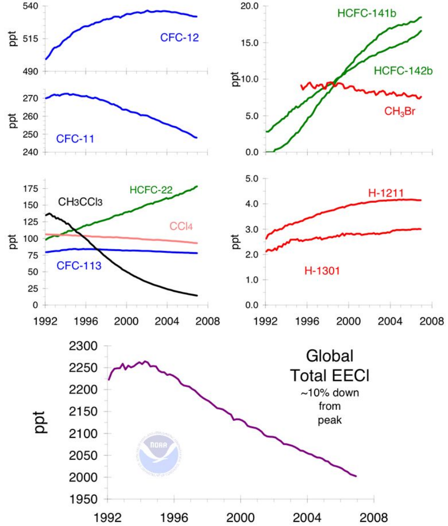
Modernity 2

- No clear culprits
  - many of us are implicated in these negative effects
  - we need modern technology to even identify and tackle risks
- The new risks are far more evenly distributed
  - Modernity 1
    - Living next to a factory was risky, if rich you could move away
  - Modernity 2
    - Ozone, global warming
    - arguably you can only temporarily can avoid these risks with your personal wealth
- **You will operate in a more regulated environment than any generation before**

Solution attempt #2
Regulation

Reasons to regulate

- Avoid harm
- Pushback
    - “regulation stifles innovation”
        - answer1: there are worse things than slowed innovation)
        - answer2: Porter’s hypothesis



Review of Environmental Economics and Policy

Volume 17 Issue 2 Summer 2023

Editorial

Editor’s Introduction, *Spencer Banzhaf*

Articles

Structural Transformation, Agriculture, Climate, and the Environment, *Christopher B. Barrett, Ariel Ortiz-Bobea, and Trinh Pham*

What Drives and Stops Deforestation, Reforestation, and Forest Degradation? An Updated Meta-analysis, *Jonah Busch and Kalifi Ferretti-Gallon*

The Economics of Natural Gas Flaring and Methane Emissions in US Shale: An Agenda for Research and Policy, *Mark Agerton, Ben Gilbert, and Gregory B. Upton Jr.*

The Economics of Electric Vehicles, *David S. Rapson and Erich Muehlegger*

The Fiscal Implications of the US Transition Away from Fossil Fuels, *Daniel Raimi, Emily Grubert, Jake Higdon, Gilbert Metcalf, Sophie Pesek, and Devyani Singh*

Invention and Global Diffusion of Technologies for Climate Change Adaptation: A Patent Analysis, *Simon Touboul, Matthieu Glachant, Antoine Dechezleprêtre, Sam Fankhauser, and Jana Stoever*

Climate Change, Epidemics, and Inequality, *Belinda Archibong and Francis Annan*

Features

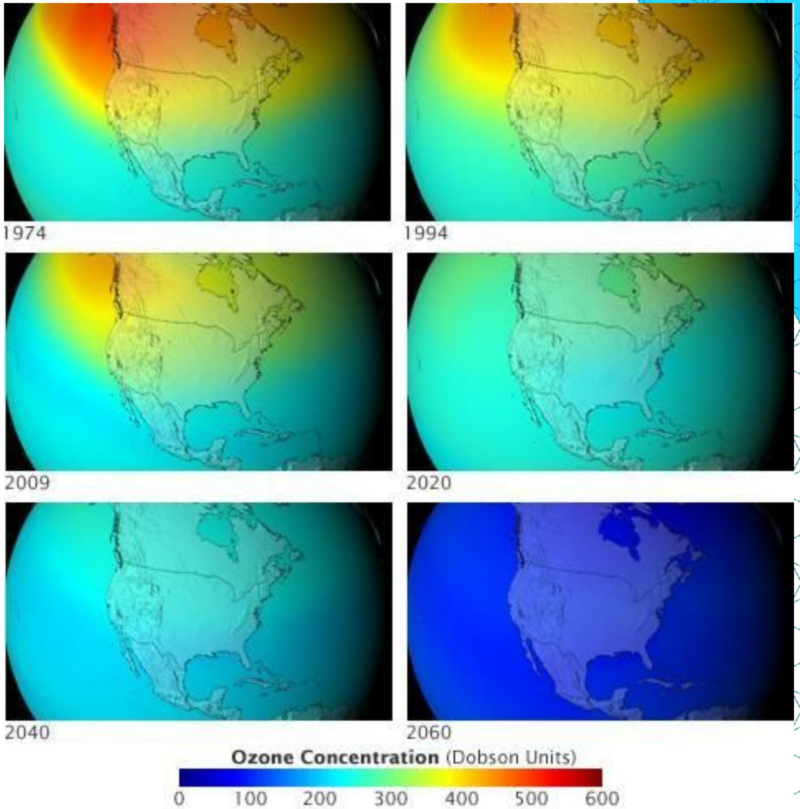
Environmental Justice Analysis for EPA Rulemakings: Opportunities and Challenges, *Ann Wolverton*

Announcements

AERE

EAERE European Association of Environmental and Resource Economists

Published by the University of Chicago Press on behalf of the Association of Environmental and Resource Economists (AERE) and the European Association of Environmental and Resource Economists (EAERE)



Chlorofluorocarbon (CFC)
Wikimedia Commons: Ozone Layer depletion if the CFC was not banned

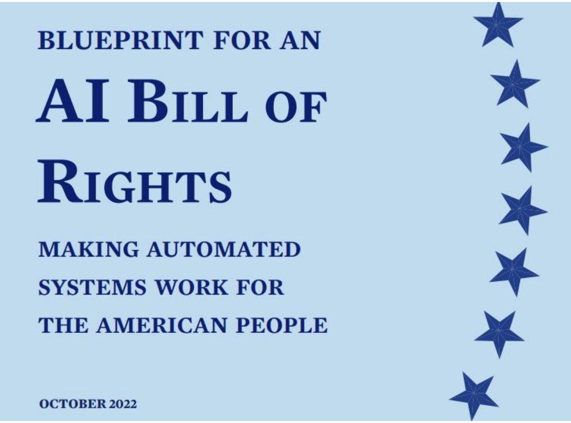
Soft Law / Recommendations

The European Commission's

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

DRAFT
ETHICS GUIDELINES
FOR TRUSTWORTHY AI

ETHICALLY ALIGNED DESIGN
First Edition



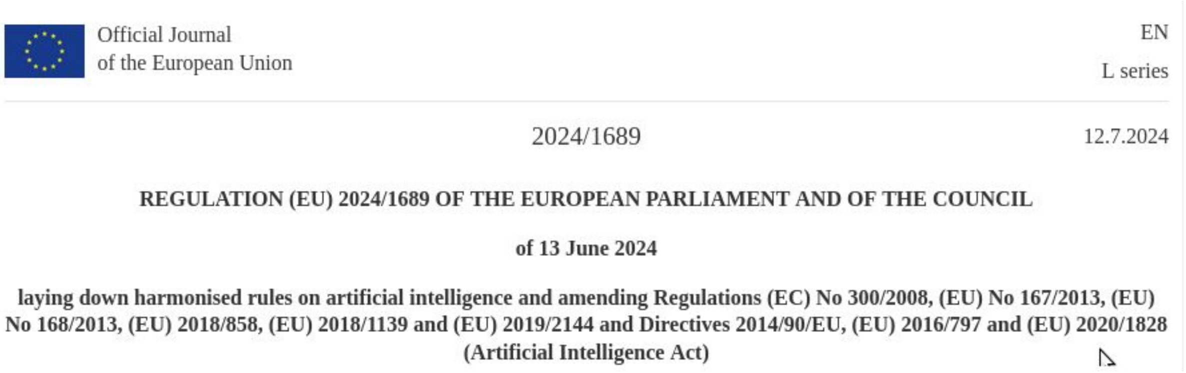
BLUEPRINT FOR AN
AI BILL OF RIGHTS

MAKING AUTOMATED
SYSTEMS WORK FOR
THE AMERICAN PEOPLE

OCTOBER 2022

Recommendations, Guidelines and Law

AI Act



Official Journal
of the European Union

EN
L series

2024/1689 12.7.2024

REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 13 June 2024

laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)

https://artificialintelligenceact.eu/assessment/eu-ai-act-compliance-checker/

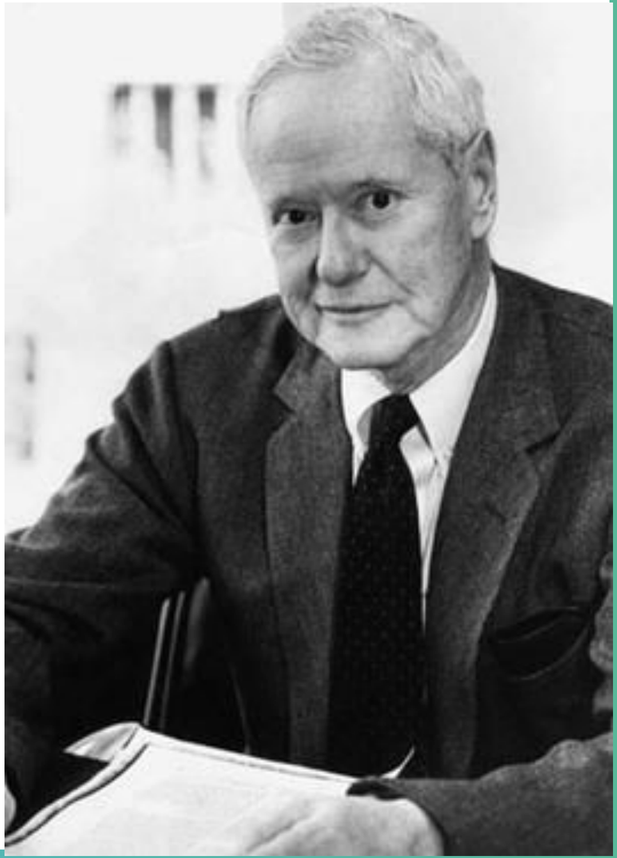
Why are we here - revisited

Why are we here: to learn

- To understand the stakes
    - **technological lock-in**
- To learn about the methods
    - **machine ethics**, guidelines, technical standards
- To understand the demands of society
    - **rules** & regulations

Why are we here: to think & challenge

- Personal Integrity
  - Challenging the
    - **customer**,
    - the **technology** and
    - the **rules**
    - **yourself**
  - Ability to walk away
- It takes knowledge of
  - history
  - argumentation
  - ethics
- CUDOS - the ethos of science
  - **C**ommon discoveries
  - **U**niversal knowledge
  - **D**isinterestedness
  - **O**rganized **S**cepticism



Robert K. Merton
source:
WikiMedia
Commons

Why are we here: to meet

- everyone needs acknowledgement
- in human-centered AI your are most likely to get this from your peers

retroreflective vests

bicycle retroreflectors

Thank you for your attention!
Questions?