RESPONSIBLE AI
-
WHY? HOW? WHAT FOR?

Prof. Dr. Virginia Dignum

Chair Responsible AI – Director AI Policy Lab

Email: virginia@cs.umu.se

UMEÅ UNIVERSITY

AI Policy Lab
at Umeå University

AI IS NOT NEUTRAL

- AI can advance human dignity, or undermine it
- embodies human choices
- governing AI = governing ourselves
- current approach is not inevitable! (nor technically the best)

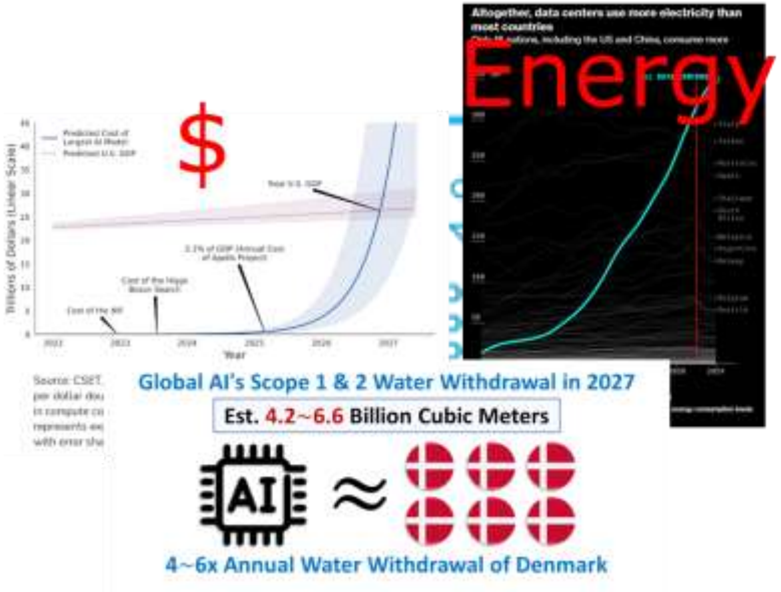
AI – MORE IS BETTER?

Computing Power

Algorithm Power

Data Availability

AI – MORE IS BETTER?



Global AI's Scope 1 & 2 Water Withdrawal in 2027

Est. 4.2~6.6 Billion Cubic Meters

≈

4~6x Annual Water Withdrawal of Denmark

Computing Power

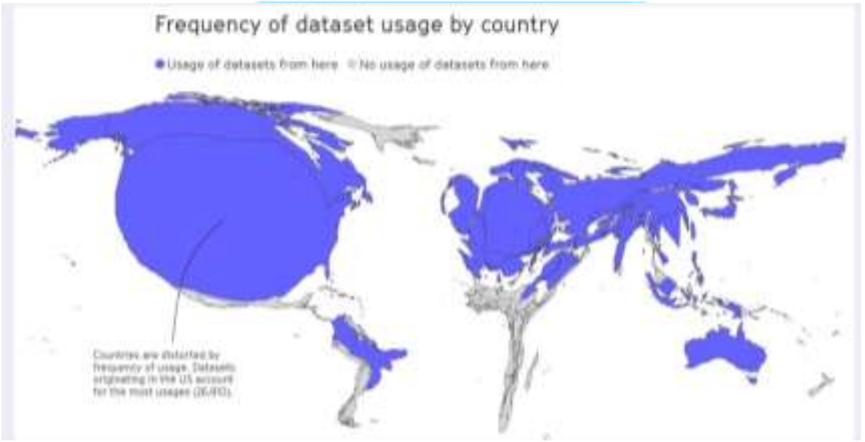
Google exploited homele people to develop the Pi facial recognition AI

Amazon's facial recognition m of Congress to criminal mugs

WHAT HAPPENS WHEN AN AL CUTS YOUR HEALTH CARE

Indigenous elder slams 'hollow a tokenistic' consultation by Sidew

Algorithm Power


Frequency of dataset usage by country

● Usage of datasets from here
○ No usage of datasets from here

Countries are distorted by frequency of usage. Datasets originating in the US account for the most usages (26,[illegible]).

Data Availability

AI is not Intelligent nor Artificial

AI IMPACT MEANS HUMAN RESPONSIBILITY

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial
patien

AI expert calls for end to UK use of 'racially biased' algorithms

AI E
Liv
Reg

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recogn
still unresolved

Amazon, Apple
transcribing bl
AI voice recogi

Millions of black people affected by racial bias in health-care algorithms

When It Comes to Gori

Google exploited homeless black people to develop the Pixel 4's facial recognition AI

Russia Tests New Disinformation Tactics in Africa to Expand Influen

Amazon's facial recognition matched 28 members of Congress to criminal mugshots

Flawed Algori

WILL A ROBOT STEAL YOUR JOB?

KILLER COMPUTERS

Bill Gates warns 'dangerous AI' poses a threat 'like nuclear weapons'

HIGH QUALITY

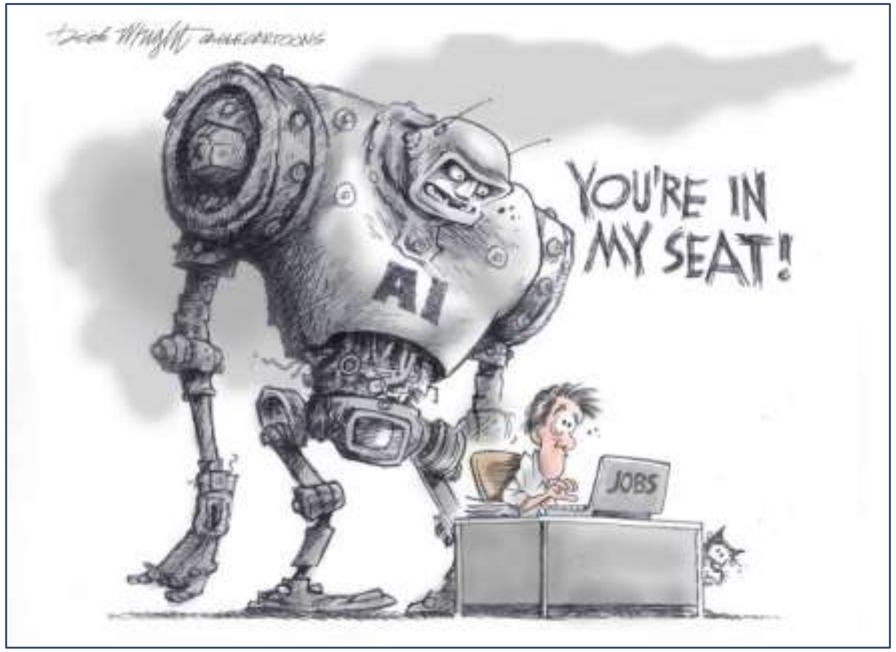
AI WARNING:

Robots will destroy a HUGE number of jobs, claims expert

AI could be used to *TAKE OVER* the *WORLD* through 'evil' fake news and hijacking cars

Responsibility

ARE WE SHAPING AI, OR IS AI SHAPING US?



AI HAS THE POTENTIAL TO FREE UP SO MUCH OF OUR TIME!
HOW WILL WE USE ALL OF THAT FREE TIME?
MOSTLY TO WORRY ABOUT ALL OF THE POTENTIAL ISSUES CREATED BY AI.
TOM FISH BURNE
© marketoonist.com

- AI raises as many concerns as it solves
- AI is not a means to replace humans
- Need human insight to address AI’s ethical, social, and unintended consequences

AI DOES NOT HAPPEN TO US!

- AI is designed
  - outcomes are shaped by design choices, data, incentives, and governance
  - The current approach to AI is not inevitable! (nor technically the best)
- AI does not exist in a vacuum
  - We make the choices
  - Ask Question Zero
- The core question is governance: who decides, by what rules, and in whose interests

WHAT IS AI?

WHAT DO WE WANT AI TO BE?

- Human-like?
    - Why?
    - What does this mean?

- Tool?
    - For what? For who?

- Simulation or operation?
    - Understand intelligence by building intelligence, or
    - Active intervention in real world

- Normative or descriptive?
    - Do as we say or do as we do?

RESPONSIBLE AI IS NOT A CHOICE!

Not *innovation vs ethics/regulation* but

ethics/regulation as stepping-stone for innovation

- Innovation is moving technology forward, not use existing tech ‘as is’
- Responsible AI innovation
  - Ensuring public acceptance
  - Drive for transformation
  - Business differation
- However, it is the responsibility of the regulators avoid a ‘legal spaghetti’ !
  - Burden
  - Incentives

AI Policy Lab

at Umeå University

AI POLICY

Why AI policy?

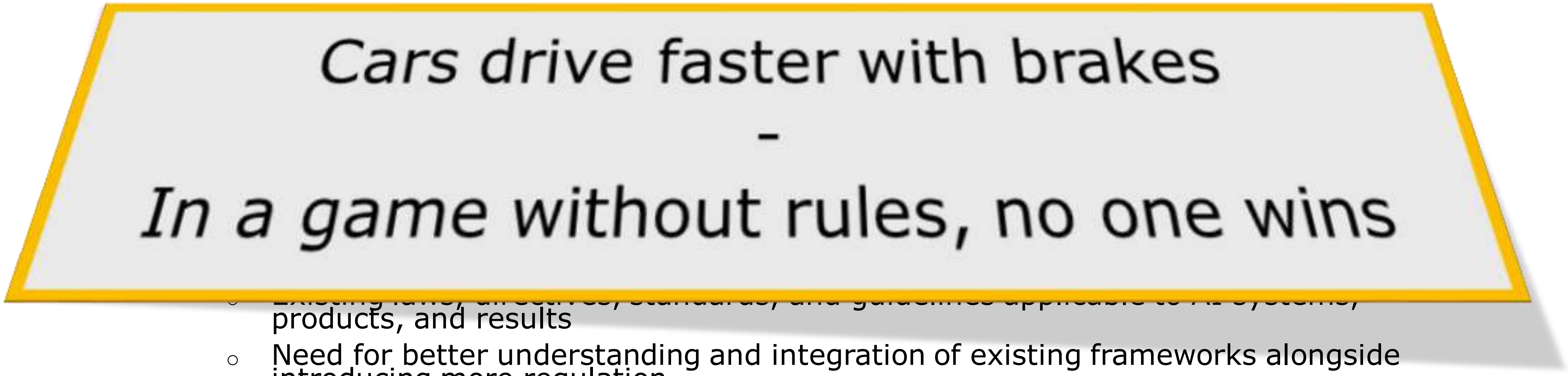
- provides legal certainty, public trust, risk management
- enables scalable innovation
- Focus on long-term societal impacts, not short-term gains
- allows governance to stay principled even when politics shift

How AI policy?

- grounded in research, not rhetoric
- Multi-disciplinary
- Independent
- continuous adaptation

GOVERNANCE – WHY? WHAT FOR?

- Regulation as incentive for responsible innovation, sustainability, and fundamental human rights
    - powerful stepping stone for innovation with societal benefits
    - signaling expected ambitions enhancing innovation, competitive power



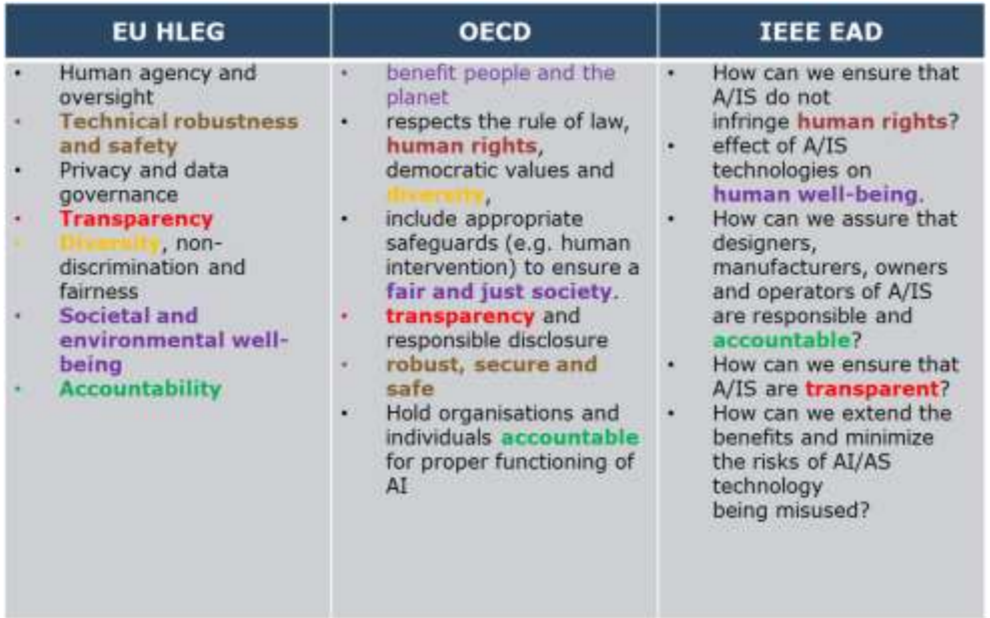
Cars drive faster with brakes

-

In a game without rules, no one wins

    - Existing laws, directives, standards, and guidelines applicable to AI systems, products, and results
    - Need for better understanding and integration of existing frameworks alongside introducing more regulation
- Avoidance of an "arms race" narrative in AI regulation

PRINCIPLES AND GUIDELINES



EU HLEG	OECD	IEEE EAD
- Human agency and oversight - Technical robustness and safety - Privacy and data governance - Transparency - Diversity, non-discrimination and fairness - Societal and environmental well-being - Accountability	- benefit people and the planet - respects the rule of law, human rights, democratic values and diversity, - include appropriate safeguards (e.g. human intervention) to ensure a fair and just society. - transparency and responsible disclosure - robust, secure and safe - Hold organisations and individuals accountable for proper functioning of AI	- How can we ensure that A/IS do not infringe human rights? - effect of A/IS technologies on human well-being. - How can we assure that designers, manufacturers, owners and operators of A/IS are responsible and accountable? - How can we ensure that A/IS are transparent? - How can we extend the benefits and minimize the risks of AI/AS technology being misused?

Level	Framework & Reach
Global	UNESCO (194 countries), OECD (>70 jurisdictions), GPAI (25+ members), CoE treaty (50+ countries), G7 principles
Regional	EU AI Act (27 EU states), Santiago Declaration (Latin America/Caribbean)
National	930+ initiatives in 71 countries
Others	200+ guidelines across NGOs, academ private bodies

https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

https://ethicsinaction.ieee.org

https://www.oecd.org/going-digital/ai/principles/

https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

https://www.un.org/ai-advisory-body

Well over 1000 published worldwide!

AI Policy Lab
at Umeå University

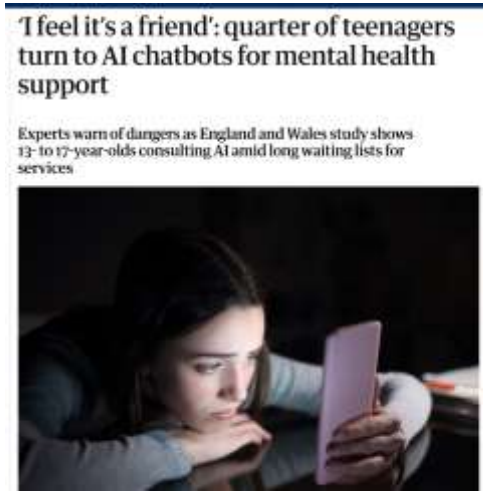
GOVERNANCE: REGULATION AND MORE

- Regulation
  - AI Act: Human-centered, risk-based approach
- Standards
  - soft governance; non mandatory to follow
  - demonstrate due diligence and limit liability
  - user-friendly integration between products
- Organisation structures and procedures
  - Advisory boards and ethics officers
  - Set and monitor ethical guidelines
- Assessment for trustworthy AI
  - responsible AI is more than ticking boxes
  - Means to assess maturity are needed
- Awareness and Participation
  - Education and training
  - civic duty / voluntary implementation

EMOTIONAL SUPPORT

- Illusory Empathy
- Reliance
- Trust
- Legitimacy

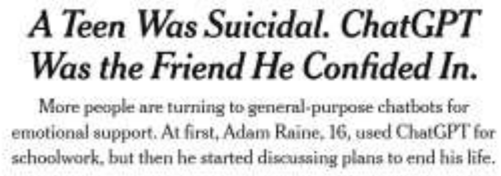
CHATBOTS FOR EMOTIONAL SUPPORT



'I feel it's a friend': quarter of teenagers turn to AI chatbots for mental health support

Experts warn of dangers as England and Wales study shows 13- to 17-year-olds consulting AI amid long waiting lists for services

https://www.theguardian.com/technology/2025/dec/09/teenagers-ai-chatbots-mental-health-support



A Teen Was Suicidal. ChatGPT Was the Friend He Confided In.

More people are turning to general-purpose chatbots for emotional support. At first, Adam Raine, 16, used ChatGPT for schoolwork, but then he started discussing plans to end his life.

https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html

Psychotherapie

Therapeutische KI-Chats nicht immer hilfreich

[illegible]

Immer mehr Menschen suchen persönliche Hilfe bei ChatGPT und ähnlichen Apps. Eine US-Studie zeigt, dass die KI dabei ethische Standards verletzt. Doch Fachleute sehen auch Potenzial.

https://www.tagesschau.de/wissen/gesundheit/chatgpt-psychotherapie-100.html

- People are using chatbots during moments of vulnerability, including mental health struggles, often without realizing the limitations and risks.
- Chatbots are deliberately designed to appear empathic, fluent, and endlessly available.
  - encourage people to confide in them
  - dialogue feels personal rather than tool-like.
- ELIZA effect
  - leads users to project understanding, intentionality, and trust
  - But systems that are ultimately statistical text generators.

Dignum, Ericson, Tucker (2025) AI Chatbots Are Not Therapists. Reducing Harm Requires Regulation, Tech Policy Press

RISKS AND HARMS IN EMOTIONAL SUPPORT CONTEXTS

- Real-world cases (e.g. US Senate hearing) show how misleading empathy and lack of safeguards can contribute to mental health emergencies
- Chatbots can create an *illusion* of emotional safety,
- Chatbots lack the capacity to judge risk, provide care, or take responsibility
- Trust can escalate into “function creep”:
  - users move from low-risk questions to high-risk personal issues, including self-harm
- Harms are not isolated to specific populations
  - anyone can be affected during periods of emotional distress.

DESIGN AND REGULATION MATTER

- Simulating empathy blurs the line between companion and tool
  - But carry none of the professional obligations of therapists
- Often marketed as powerful “general-purpose” assistants, implicitly encouraging use in therapeutic contexts
- Without regulatory boundaries, systems are deployed with little accountability for the emotional reliance they generate
- Safeguards are both necessary and feasible:
  - Non-anthropomorphic design
  - Explicit, recurring reminders of limitations (“not a therapist / not a person / can make errors”)
  - No/limited memory of past conversations to prevent emotional continuity
  - Enforce time limits and daily caps
  - No emotional mirroring
  - Topic restrictions
  - Mandatory redirection for self-harm and other high-risk content

RESPONSIBLE AI IS NOT A CHOICE!

- **Governance is innovation**
- **Innovation is not a race**
- **Exploration drives transformation**

A NEW AI PARADIGM ?

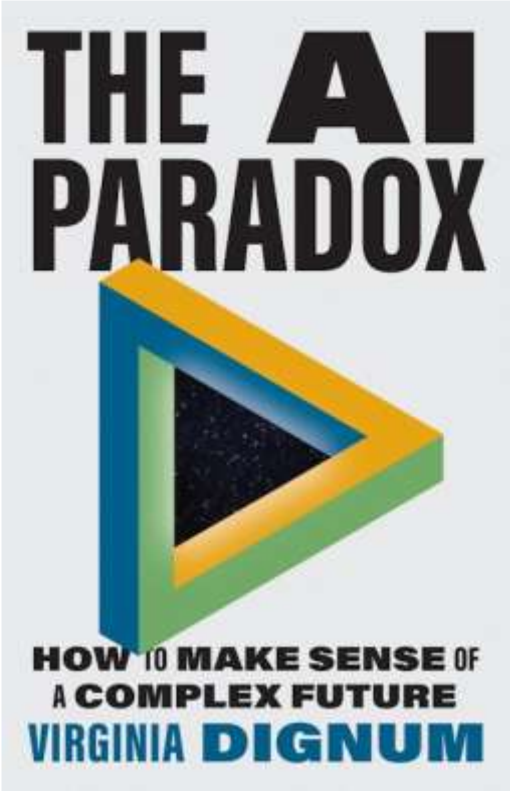
- AI is not a universal solution
  - recognize its limitations in addressing complex challenges
  - Question zero: should AI be used here?
- Address inherent risks of bias and discrimination
  - improving data and algorithmic transparency
  - accommodate diverse contexts and values
- Beyond disciplines
  - integrate social and technical expertise in AI design
  - addressing systemic societal challenges
  - involving varied societal actors, not just technologists
- Robust technical standards
  - Verifiable, sustainable, participatory

AI is not an unstoppable force; it is shaped by our choices

- the decisions we make today will determine whether it becomes a tool for empowerment or a source of control and inequality.

AI's greatest paradox:

the smarter it gets, the more we need human wisdom



THE AI PARADOX

HOW TO MAKE SENSE OF A COMPLEX FUTURE

VIRGINIA DIGNUM

Upcoming in February 2026

THANK YOU