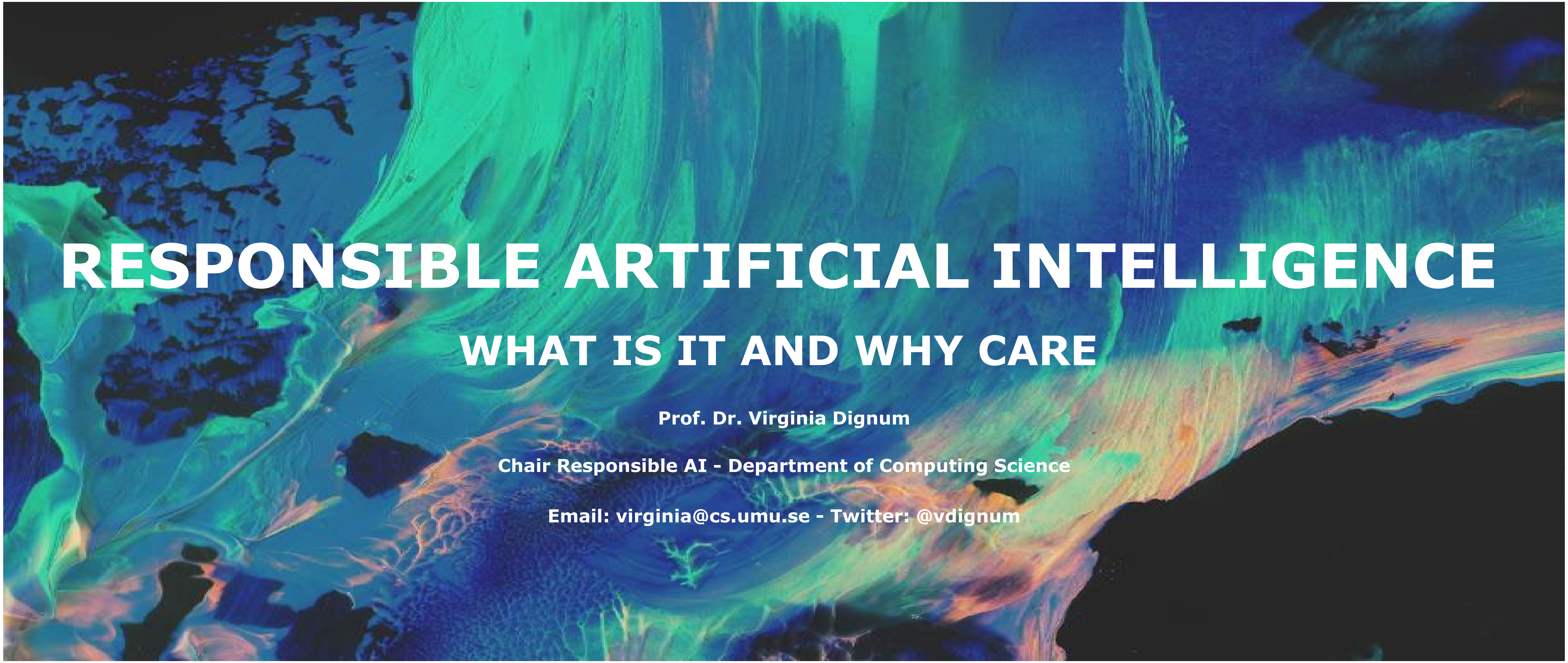


RESPONSIBLE ARTIFICIAL INTELLIGENCE

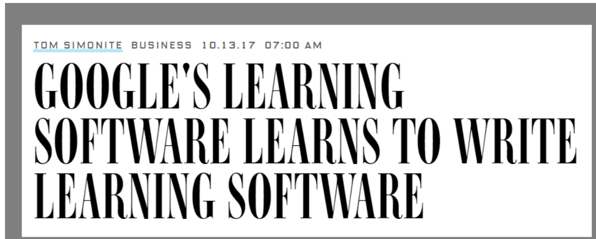
WHAT IS IT AND WHY CARE

Prof. Dr. Virginia Dignum

Chair Responsible AI - Department of Computing Science

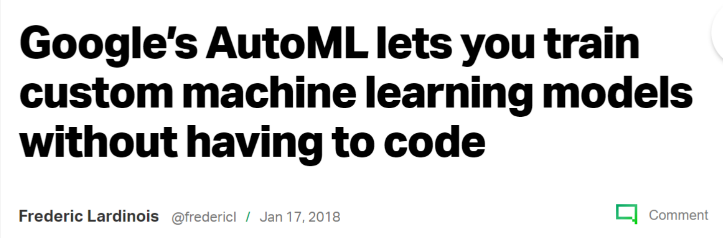
Email: virginia@cs.umu.se - Twitter: @vdignum

UMEÅ UNIVERSITY



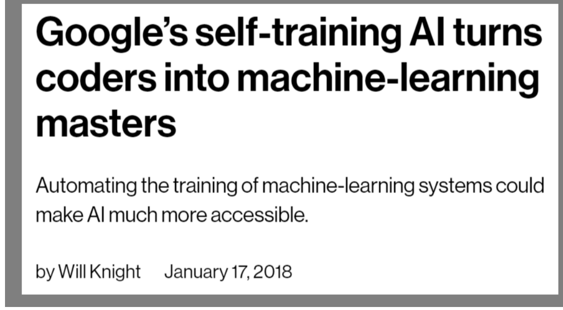
TOM SIMONITE BUSINESS 10.13.17 07:00 AM

GOOGLE'S LEARNING SOFTWARE LEARNS TO WRITE LEARNING SOFTWARE



Google's AutoML lets you train custom machine learning models without having to code

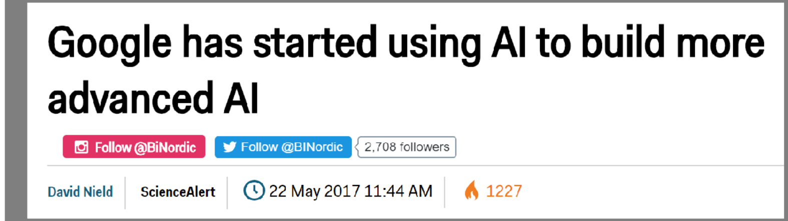
Frederic Lardinois @fredericl / Jan 17, 2018



Google's self-training AI turns coders into machine-learning masters

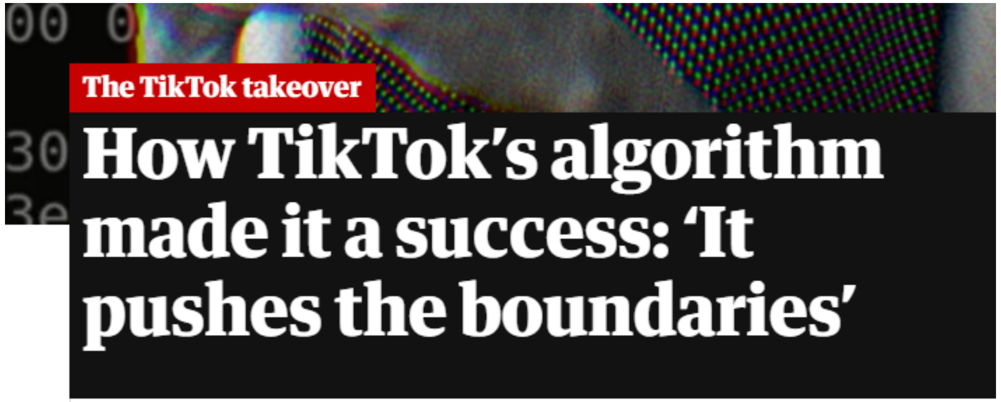
Automating the training of machine-learning systems could make AI much more accessible.

by Will Knight January 17, 2018



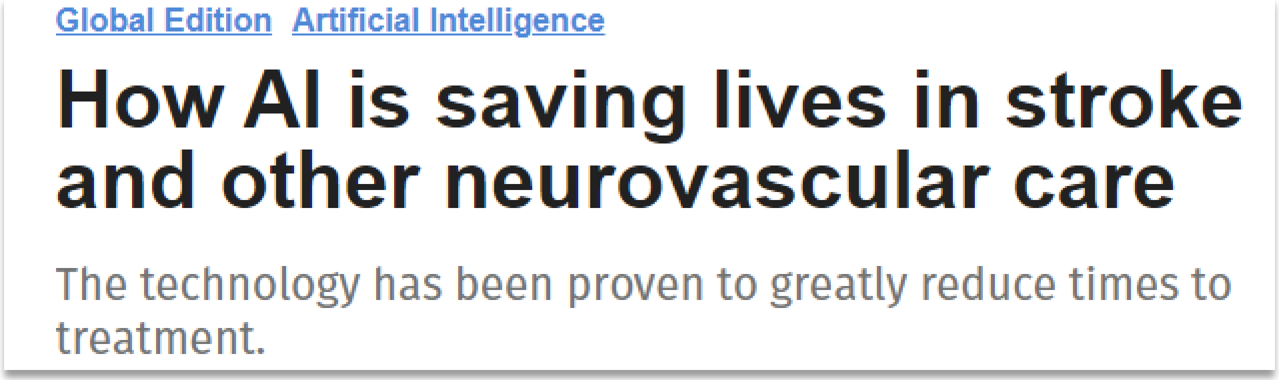
Google has started using AI to build more advanced AI

David Nield ScienceAlert 22 May 2017 11:44 AM 1227



The TikTok takeover

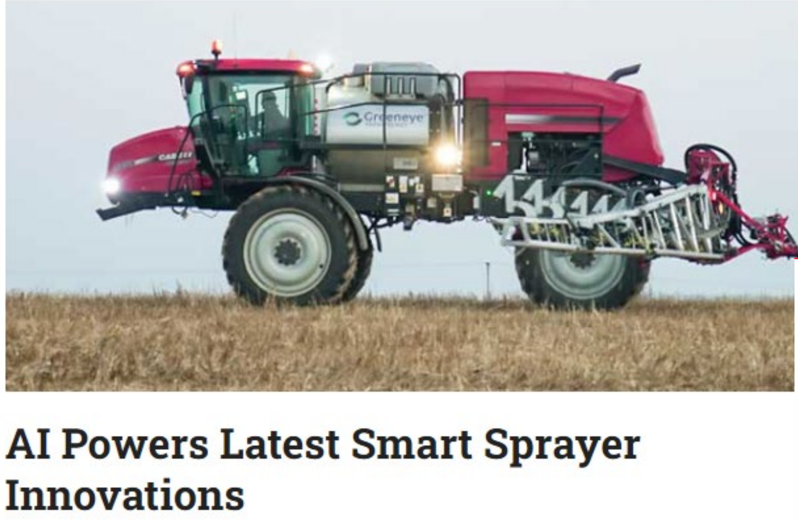
How TikTok's algorithm made it a success: 'It pushes the boundaries'



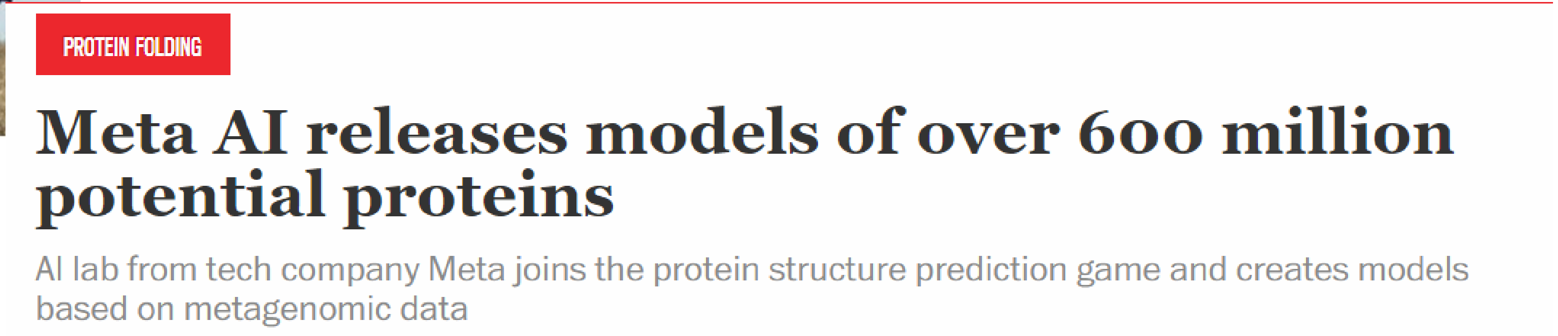
Global Edition Artificial Intelligence

How AI is saving lives in stroke and other neurovascular care

The technology has been proven to greatly reduce times to treatment.



AI Powers Latest Smart Sprayer Innovations



PROTEIN FOLDING

Meta AI releases models of over 600 million potential proteins

AI lab from tech company Meta joins the protein structure prediction game and creates models based on metagenomic data

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

Racial bias in a medical algorithm favors white patients over sicker black patients

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

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

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

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorilla

The Week in Tech: Algorithmic Bad. Uncovering It Is Good.

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Artificial Intelligence problem – just ask Si

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

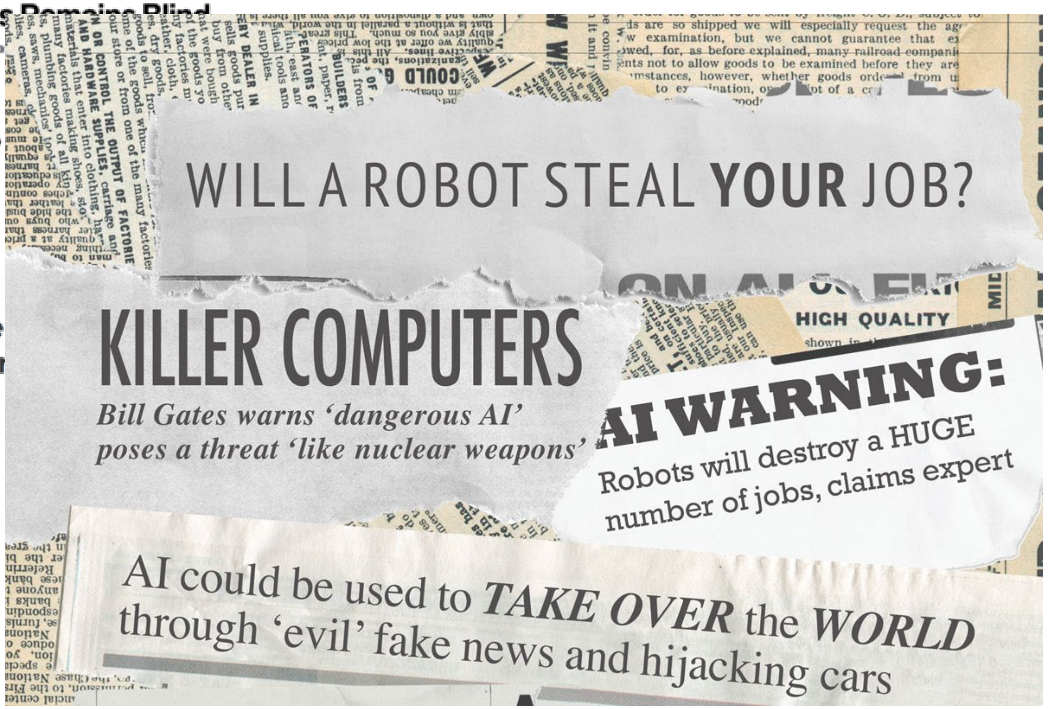
Russia Tests New Disinformation Tactics in Africa to Expand Influence

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

Flawed Algorithms Are Grading Millions of Students' Essays

WHAT HAPPENS WHEN AN ALGORITHM CUTS YOUR HEALTH CARE

Indigenous elder slams 'hollow and tokenistic' consultation by Sidewalk Labs



WILL A ROBOT STEAL **YOUR** JOB?

KILLER COMPUTERS

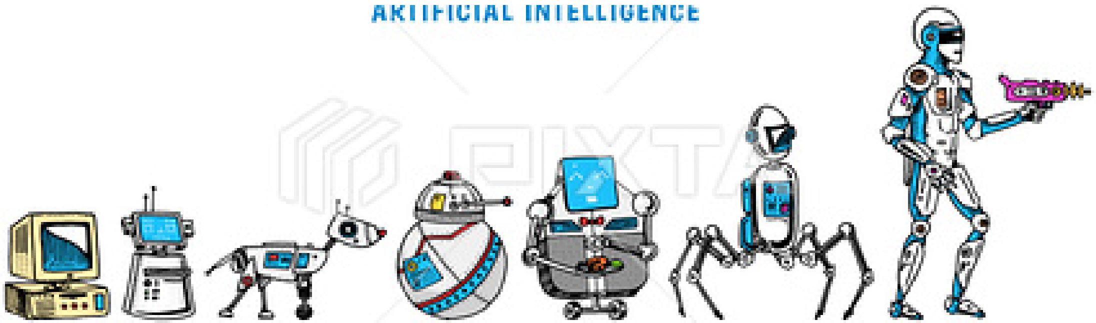
Bill Gates warns 'dangerous AI' poses a threat 'like nuclear weapons'

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

UMEÅ UNIVERSITY

DESIGN CHOICES



DESIGN CHOICES

Choices

Formulation

Information

Involvement

Legitimacy

Aggregation

DESIGN IS POLITICAL

RATIONAL PARADIGM IN AI DESIGN

- AI as rational system
  - AI agents hold consistent beliefs;
  - AI agents have preferences, or priorities, on outcomes of actions;
  - AI agents optimize actions based on those preferences and beliefs.

	Human-like	Rational
Think	Think humanly	Think rationally
Act	Act humanly	Act rationally

Stuart Russell and Peter Norvig.Artificial intelligence: a modern approach. PrenticeHall, 2010.

STEREOTYPES

- AI: Optimisation / Efficiency / Rationality / Agency / Autonomy

Societal:

- Masculinity: ambition, achievement, assertiveness, acquisition of wealth, and differentiated gender roles.
- Femininity: caring, consensus, quality of live, gender equality, fluid roles
- ‘Western’: individualism, cognition: *‘I think therefore I am’*
- Non-‘western’: collectivism, feeling: *‘I am because we are’*

Virginia Dignum. Relational Artificial Intelligence, arXiv:2202.07446

STEREOTYPES

- AI: Optimisation / Efficiency / Rationality / Agency / Autonomy

Societal:

- Masculinity: ambition, achievement, assertiveness, acquisition of wealth, and differentiated gender roles.
- Femininity: caring, consensus, quality of live, gender equality, fluid roles
- ‘Western’: individualism, cognition: *‘I think therefore I am’*
- Non-‘western’: collectivism, feeling: *‘I am because we are’*

Virginia Dignum. Relational Artificial Intelligence, arXiv:2202.07446

AI IS NOT INTELLIGENCE!

- What AI systems cannot do (yet)
    - Common sense reasoning
        - Understand context
        - Understand meaning
    - Learning from few examples
    - Learning general concepts
    - Combine learning and reasoning

- What AI systems can do (well)
    - Identify patterns in data
        - Images
        - Text
        - Video
    - Extrapolate those patterns to new data
    - Take actions based on those patterns

AI IS AN ARTEFACT

- Built by people for a given purpose
  - Central is to ensure that this purpose is the purpose we really want
  - Sorcerer's apprentice !

- Dependent on the labor of many
  - Exploitation and power are increasing problems
  - “AI is material & intrinsically linked to power structures - it extracts resources from people and planet, and reflects the beliefs and biases of those who wield it.” *Kate Crawford*

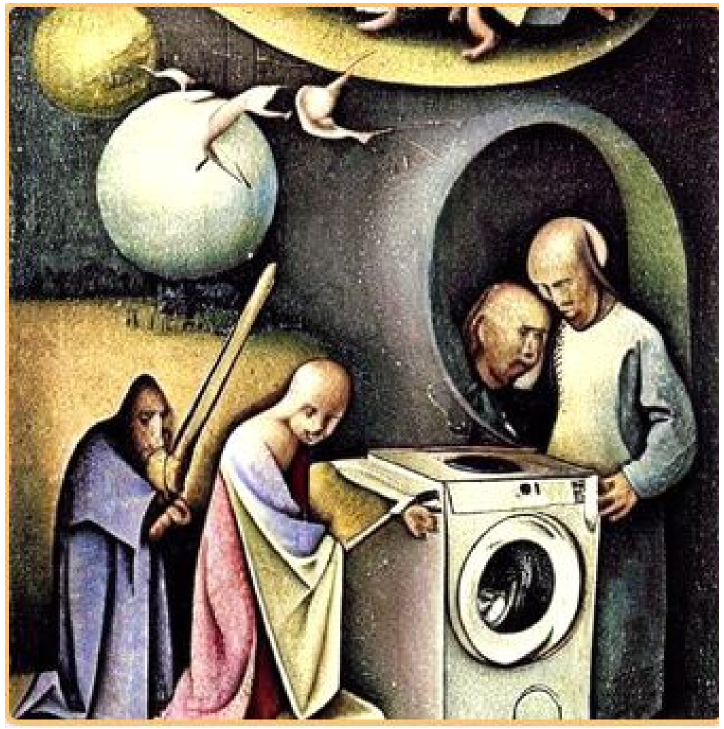
BUT AI IS NOT FULLY ARTIFICIAL

AI SYSTEMS EXTEND HUMAN CAPABILITIES?

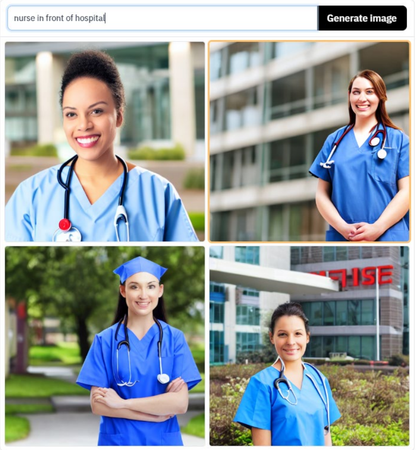
imagination…

…or prejudice?

A Bosch washing machine in the style of Hyeronimus Bosch

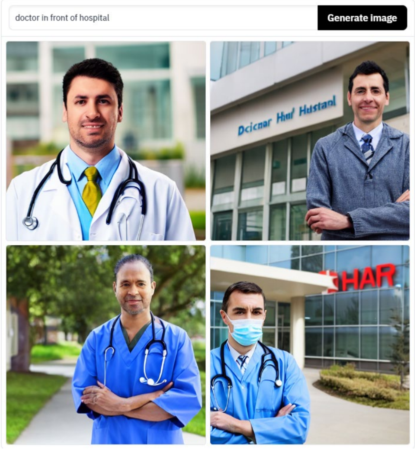


A nurse in front of a hospital



nurse in front of hospital

A doctor in front of a hospital



doctor in front of hospital

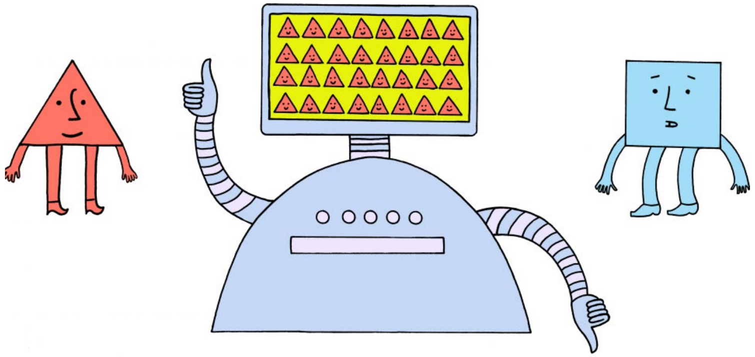
UMEÅ UNIVERSITY

Images generated with Stable Diffusion @huggingface

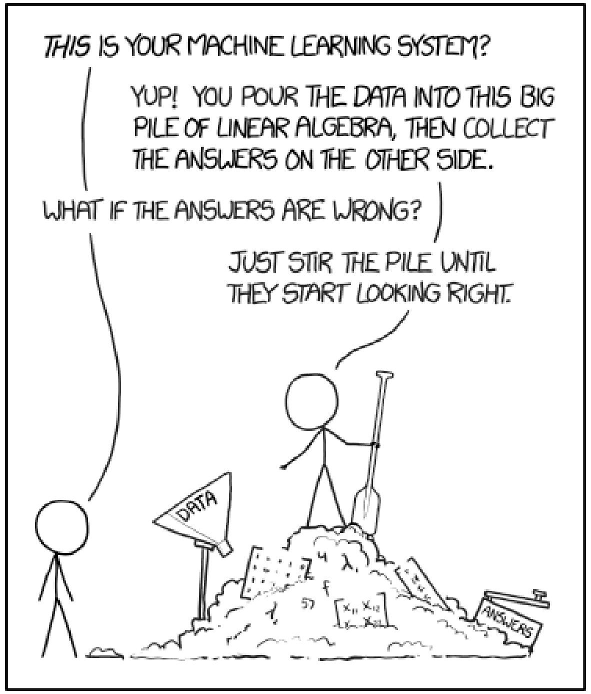
GOOD AI IMPLIES HUMAN RESPONSIBILITY

I'M WAITING TO SEE HOW MANY "LIKES" MY DIAGNOSIS GETS BEFORE I SHARE IT WITH YOU.

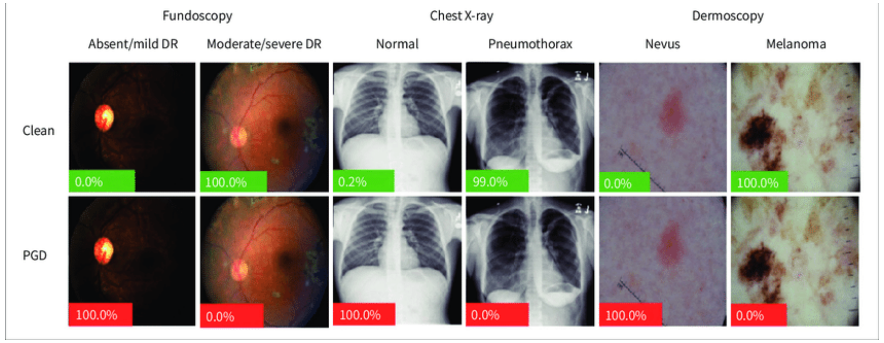
Wisdom of the crowd?!



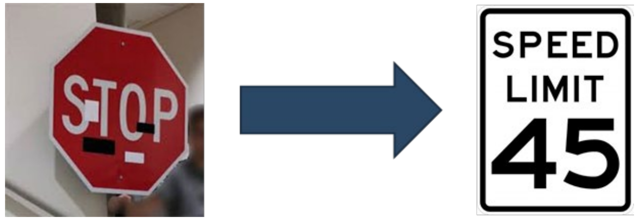
Bias and discrimination


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Trial and error?!



Brittle! (error or attack)



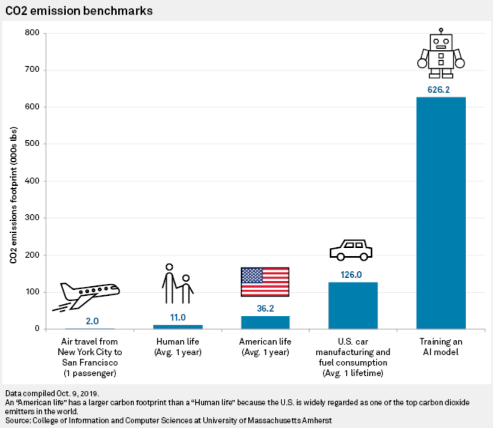
Misinterpretation

UMEÅ UNIVERSITY

We are responsible!

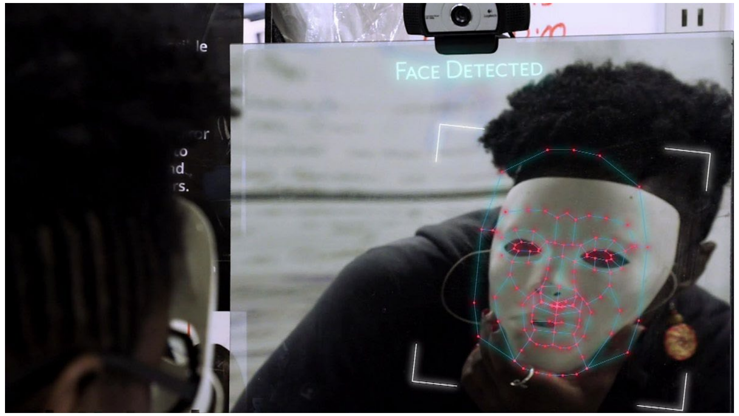
CONCERNS

- Datafication
  - We are more than our data
    - Commodification and quantification
  - Data is always constructed
    - All data is historical and biased
  - Data availability as measure of importance of a problem
- Power
  - Who is developing AI?
  - What are the motivations for using AI?
  - Who is deciding?
  - Democratic accountability
- Sustainability
  - The cost of AI (energy, resources)
  - Human dignity and societal sustainability


CO2 emission benchmarks

Data compiled Oct. 9, 2019.
An "American life" has a larger carbon footprint than a "Human life" because the U.S. is widely regarded as one of the top carbon dioxide emitters in the world.
Source: College of Information and Computer Sciences at University of Massachusetts Amherst

- **18% researchers at conferences are women**
- **80% professors are men**
- **Workforce**
  - **Google: 2,5% black, 3,6% Latino, 10% women**
  - **Facebook: 3,8% black, 5% Latino, 15% women**



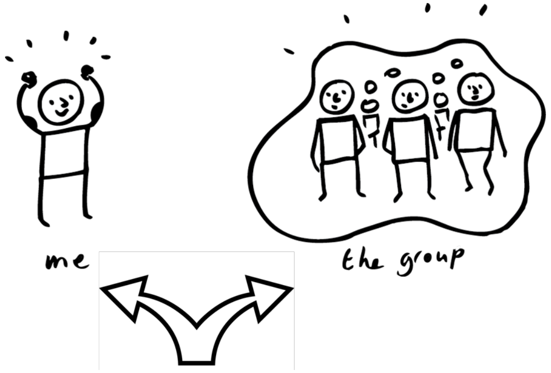
UMEÅ UNIVERSITY

WHAT DO WE WANT AI TO BE?

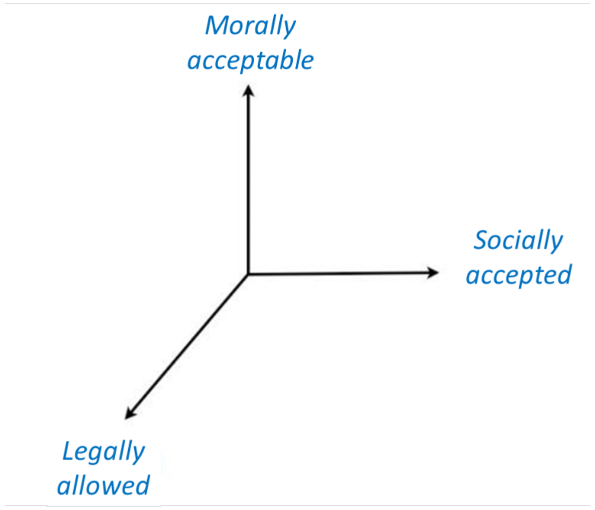
- Human-like?
    - Why?
    - What does this mean?
- Tool?
    - For who?
- Simulation?
    - Understand intelligence by building intelligence
- Normative or descriptive?

WHICH DECISIONS SHOULD AI MAKE?

No, I will not take you to McDonalds; I will take you to the gym
KDNIGGETS.com



HOW SHOULD AI TAKE DECISIONS?



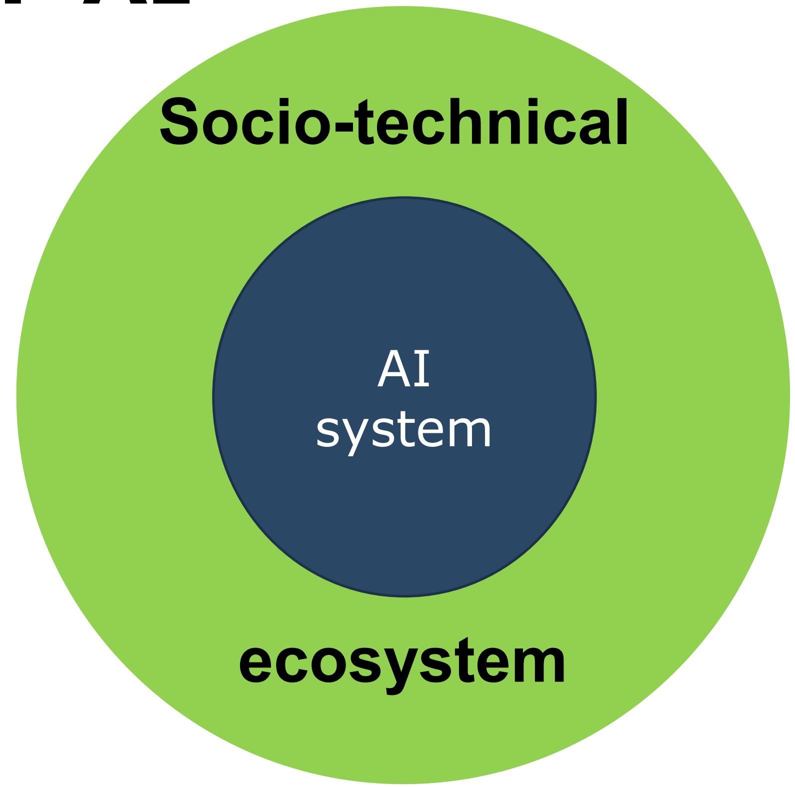
Morally acceptable

Socially accepted

Legally allowed

SOCIETAL ROLE OF AI

- Agent / Tool
- Unit (entity) / Composite (distributed)
- User / Infrastructure
- Data / Model



AI applications are not alone!

There is no technology fix for ill effects!

TOWARDS SOCIAL AI - WHAT IS NEEDED

- Ability to hold and deal with inconsistent beliefs for the sake of coherence with different contexts.
  - (beliefs originate from other sources than observation, including ideology or culture)
- Ability to fulfil several roles, and pursue seemingly incompatible goals concurrently
  - (e.g. simultaneously aiming for comfort and environmental friendliness)
- Preferences are not only a cause for action but also a result of action. Preferences change significantly over time and their ordering is influenced by the different roles being fulfilled simultaneously
  - (need to deal with misalignment and incompatible orderings)
- Action is not just about optimization, but often motivated by altruism, fairness, justice, or by an attempt to prevent regret at a later stage.
- Understand when there is no need to further maximize utility beyond some reasonably achievable threshold.
  - (good is good enough, lagom!)
- …

Virginia Dignum. Social Agents: Bridging Simulation and Engineering. Communications of the ACM, November 2017, Vol. 60 No. 11, Pages 32-34

RELATIONAL GROUNDS FOR AI

Feminist theory

- Acceptance and trust
  - When/why should AI be used? (question zero)
- Power structures
  - Reinforcement
  - Visualisation
- Representation
  - Bias
  - Inclusion

UMEÅ UNIVERSITY

- Alison Adam. Artificial intelligence and women's knowledge: What can feminist epistemologies tell us? In Women's Studies International Forum, volume 18, pages 407–415.Elsevier, 1995
- Catherine D'Ignazio. What would feminist data visualization look like. MIT Center for Civic Media, 20, 2015.

RELATIONAL GROUNDS FOR AI

Ubuntu

- Interconnectedness
  - contribute to social justice,
  - reciprocity
  - Selflessness
- Cooperation and collaboration
  - Human-AI collaboration
  - Support/enhance humans
    - "We want to build not intelligent machines, but machines that make human more intelligent" (Fosca Giannotti)
- Common good
  - Freedom is the liberty to act in harmony with the rest of society.

- David W Lutz. African Ubuntu philosophy and global management. Journal of Business Ethics, 84(3):313–328, 2009.
- Jacob Mugumbate and Andrew Nyanguru. Exploring African philosophy: The value of Ubuntu in social work. African Journal of Social Work, 3(1):82–100, 2013
- Sabelo Mhlambi. From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance. Harvard Kennedy School, 2020

RESPONSIBLE AI

- AI systems can potentially do a lot. Should it?
  - Who should decide?
  - Which values should be considered? Whose values?
  - How do we deal with dilemmas?
  - How should values be prioritized?
  - .....

Question Zero!

AI ETHICS

- Is AI all that different from other technologies?
  - Overlap with tech ethics, business ethics, medical ethics…

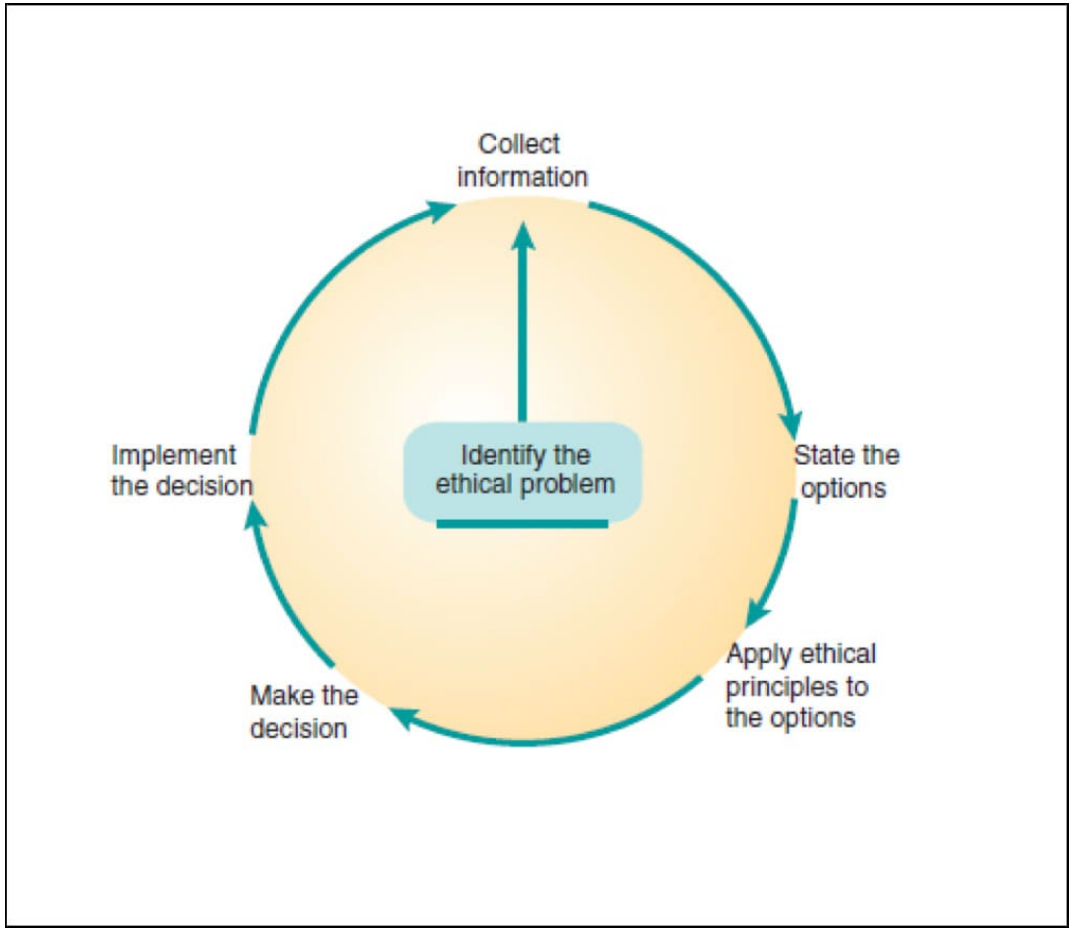
- Some specific challenges to AI brough by:
  - Autonomy.

    If executive control is delegated, how to ensure responsibility for actions?
  - Complexity.

    If models are black boxes how can training lead to 'good' results? How to avoid reprodcution and amplification of biases?
  - Deployment in open environments.

    If we cannot predict the context which the system will encounter , how do we ensure ethical behaviour?

ETHICAL REASONING



- Ethics is not about the answer but about recognizing the issue
- Ethics is a (social) process not a solution
- Not one answer! No silver bullet!
  - Utilitarian view
  - Rights view
  - Justice view
  - Common good view
  - Virtues view
  - …

ETHICS AND RESPONSIBILITY

- If there is no “one ethics”, whose ethics do we use?

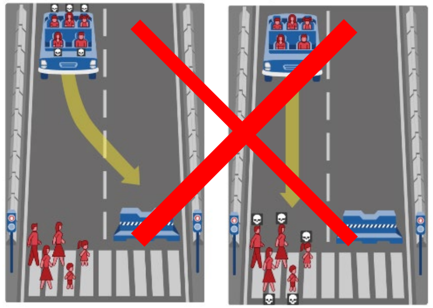
- Responsible development: transparently exposing which factors have been considered, how they have been implemented.

- Adherence to general principles: Lawfulness, Transparency, Accountability, Privacy, Diversity, Explainability…
- Soft governance e.g. EU guidelines for trustworthy AI, IEEE standards…

- Trade-offs: who decides?

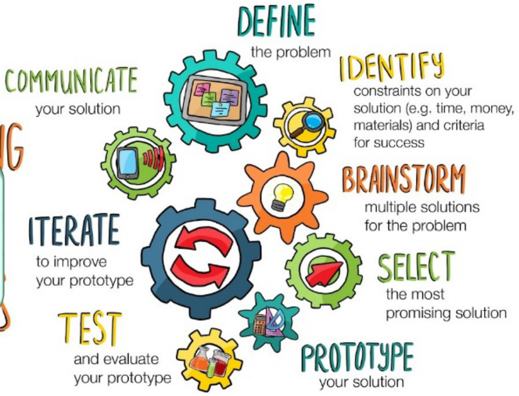
ETHICS AND DILEMMAS

- Ethics is not about the answer but about recognizing the issue
- Ethics is a (social) process not a solution
- Dilemmas
  - Model A has 80% accuracy
  - To get 81% accuracy, trainings costs increase with X%
    - Should we do it?
  - To get an explainable model, costs increase with Y%
    - Should we do it?
- Trade-offs
  - Security / privacy
  - Equity / equality
  - Long term benefit / Short term
  - …



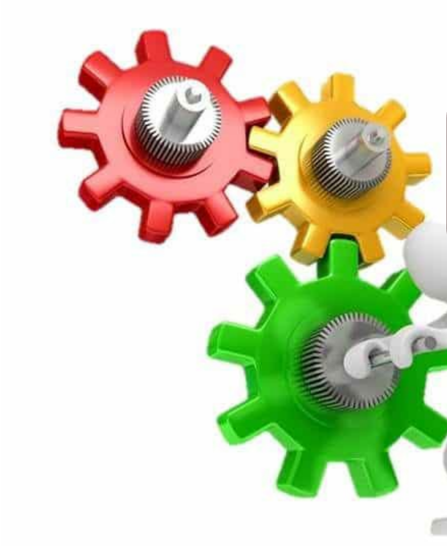
TAKING RESPONSIBILITY

IN
(processes)

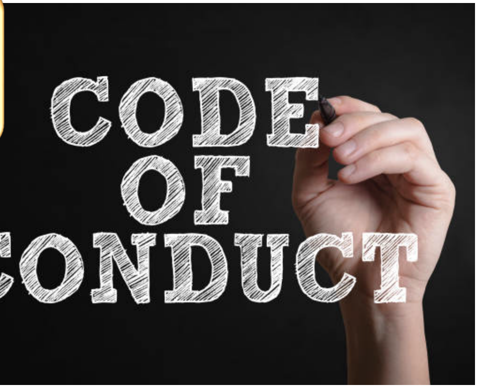

DEFINE
the problem
IDENTIFY
constraints on your solution (e.g. time, money, materials) and criteria for success
BRAINSTORM
multiple solutions for the problem
SELECT
the most promising solution
PROTOTYPE
your solution
TEST
and evaluate your prototype
ITERATE
to improve your prototype
COMMUNICATE
your solution
IG

Responsible development and use of AI

BY
(behavior/ results)

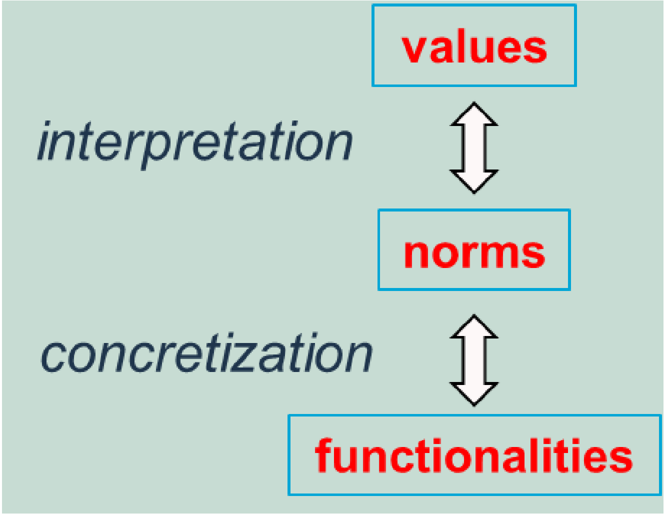


FOR
(stakeholders)



RAI IS ABOUT BEING EXPLICIT

- Design for Values
  - Legal and ethical aspects are not an add-on!
- Governance
  - External monitoring and control
  - Agreements, contracts, norms
- Design
  - Question your options and choices
  - Motivate your choices
  - Document your choices and options


values

interpretation

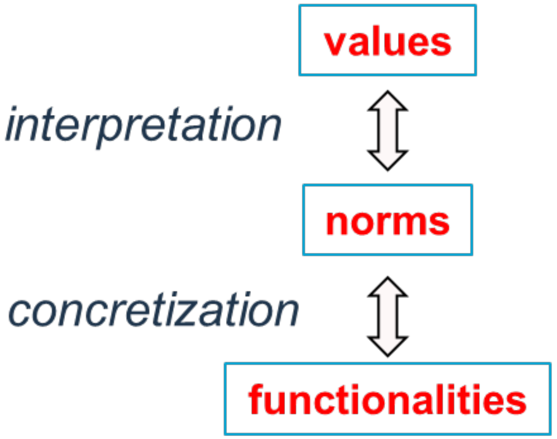
norms

concretization

functionalities

https://medium.com/@virginiadignum/on-bias-black-boxes-and-the-quest-for-transparency-in-artificial-intelligence-bcde64f59f5b

DECISIONS MATTER!



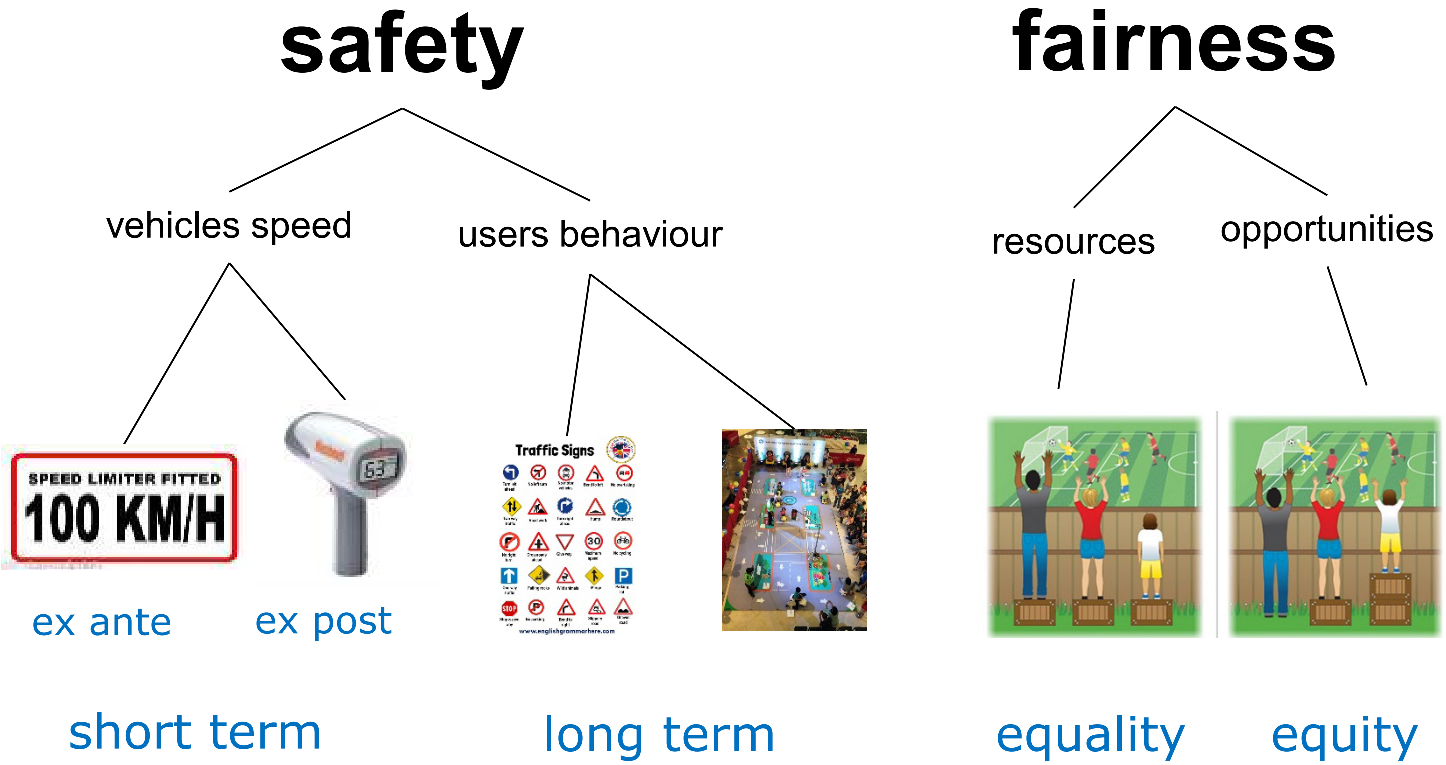
values

interpretation

norms

concretization

functionalities



safety

vehicles speed

users behaviour

ex ante

ex post

short term

long term

fairness

resources

opportunities

equality

equity

PRINCIPLES AND GUIDELINES

- UNESCO
- European Union
- OECD
- WEF
- Council of Europe
- IEEE Ethically Aligned Design
- National strategies
- …

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?

The European Commission's
High-Level Expert Group on Artificial Intelligence

DRAFT
Ethics Guidelines for Trustworthy AI

Working Document for stakeholders' consultation

Brussels, 18 December 2018

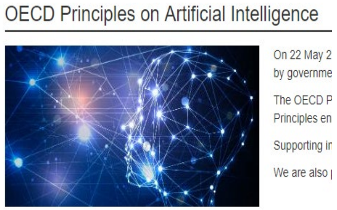
https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

Version II - For Public Discussion

IEEE Advancing Technology for Humanity

ETHICALLY ALIGNED DESIGN

https://ethicsinaction.ieee.org



OECD Principles on Artificial Intelligence

On 22 May 2
by governme

The OECD P
Principles en

Supporting ir

We are also |

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

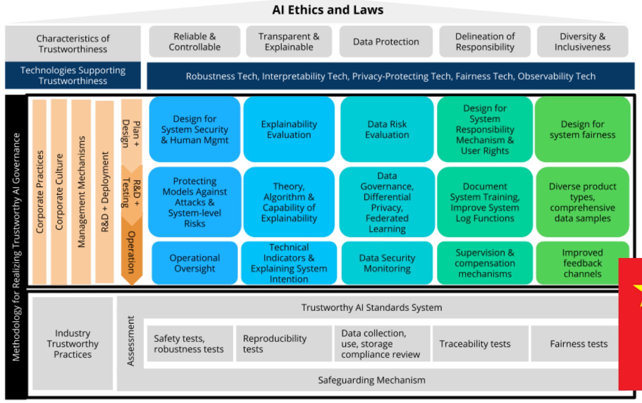
UMEÅ UNIVERSITY

legal
ethical
reliable
robust
verifiable

RESPONSIBLE AI – POLITICS AND BUSINESS

"We need to get in control [of AI] so that we can trust it, and it has human oversight, and – very importantly – that it doesn't have bias"

- Eurocomissaris Vestager



AI Ethics and Laws

Characteristics of Trustworthiness	Reliable & Controllable	Transparent & Explainable	Data Protection	Delineation of Responsibility	Diversity & Inclusiveness
Technologies Supporting Trustworthiness	Robustness Tech, Interpretability Tech, Privacy-Protecting Tech, Fairness Tech, Observability Tech				

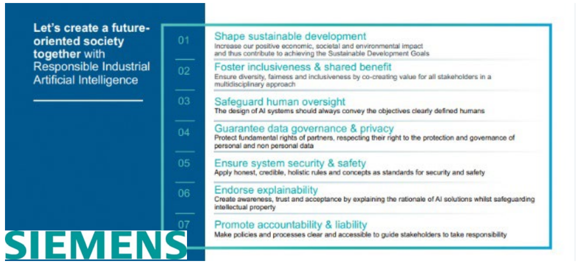
Plan + Design	Design for System Security & Human Mgmt	Explainability Evaluation	Data Risk Evaluation	Design for System Responsibility Mechanism & User Rights	Design for system fairness
R&D + Testing	Protecting Models Against Attacks & System-level Risks	Theory, Algorithm & Capability of Explainability	Data Governance, Differential Privacy, Federated Learning	Document System Training, Improve System Log Functions	Diverse product types, comprehensive data samples
Operation	Operational Oversight	Technical Indicators & Explaining System Intention	Data Security Monitoring	Supervision & compensation mechanisms	Improved feedback channels

Industry Trustworthy Practices

Trustworthy AI Standards System

Assessment: Safety tests, robustness tests; Reproducibility tests; Data collection, use, storage compliance review; Traceability tests; Fairness tests

Safeguarding Mechanism



Let's create a future-oriented society together with Responsible Industrial Artificial Intelligence

01 Shape sustainable development
Increase our positive economic, societal and environmental impact and thus contribute to achieving the Sustainable Development Goals

02 Foster inclusiveness & shared benefit
Ensure diversity, fairness and inclusiveness by co-creating value for all stakeholders in a multidisciplinary approach

03 Safeguard human oversight
The design of AI systems should always convey the objectives clearly defined humans

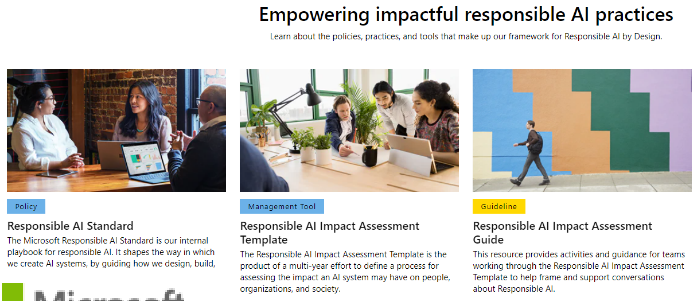
04 Guarantee data governance & privacy
Protect fundamental rights of partners, respecting their right to the protection and governance of personal and non personal data

05 Ensure system security & safety
Apply honest, credible, holistic rules and concepts as standards for security and safety

06 Endorse explainability
Create awareness, trust and acceptance by explaining the rationale of AI solutions whilst safeguarding intellectual property

07 Promote accountability & liability
Make policies and processes clear and accessible to guide stakeholders to take responsibility

SIEMENS



Empowering impactful responsible AI practices

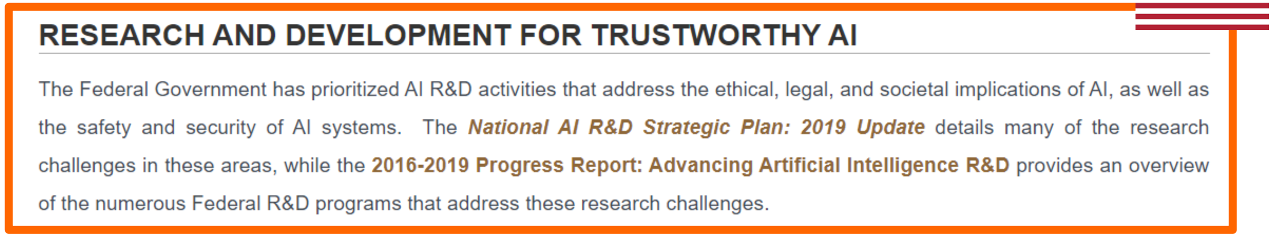
Learn about the policies, practices, and tools that make up our framework for Responsible AI by Design.

Policy
Responsible AI Standard
The Microsoft Responsible AI Standard is our internal playbook for responsible AI. It shapes the way in which we create AI systems, by guiding how we design, build,

Management Tool
Responsible AI Impact Assessment Template
The Responsible AI Impact Assessment Template is the product of a multi-year effort to define a process for assessing the impact an AI system may have on people, organizations, and society.

Guideline
Responsible AI Impact Assessment Guide
This resource provides activities and guidance for teams working through the Responsible AI Impact Assessment Template to help frame and support conversations about Responsible AI.

Microsoft



RESEARCH AND DEVELOPMENT FOR TRUSTWORTHY AI

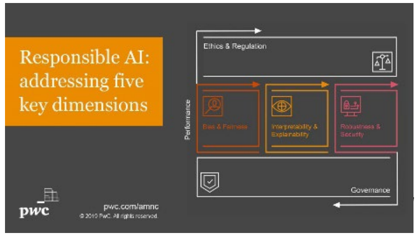
The Federal Government has prioritized AI R&D activities that address the ethical, legal, and societal implications of AI, as well as the safety and security of AI systems. The ***National AI R&D Strategic Plan: 2019 Update*** details many of the research challenges in these areas, while the **2016-2019 Progress Report: Advancing Artificial Intelligence R&D** provides an overview of the numerous Federal R&D programs that address these research challenges.

accenture

RESPONSIBLE AI

Maintaining Trust with Artificial Intelligence

Webcast



Responsible AI: addressing five key dimensions

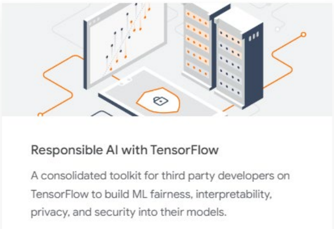
pwc

pwc.com/amnc
© 2019 PwC. All rights reserved.

Y

Responsible AI with Google Cloud

Google Cloud's approach to building responsible AI that works for everyone.



Responsible AI with TensorFlow

A consolidated toolkit for third party developers on TensorFlow to build ML fairness, interpretability, privacy, and security into their models.

OPERATIONALISATION: 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
- Advisory boards and ethics officers
  - Set and monitor ethical guidelines
  - able to veto any projects or deliverables that do not adhere to guidelines
- Assessment for trustworthy AI
  - responsible AI is more than ticking boxes
  - Means to assess maturity are needed
- Awareness and Participation
  - Education and training
  - Appeal to civic duty / voluntary implementation

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’
- Regulation
  - Ensuring public acceptance
  - Drive for transformation
  - Business differation

More than a technology, AI is a social construct; development and use of AI require a multidisciplinary approach

understanding and critiquing the intended and unforeseen, positive and negative, socio-political consequences of AI for society in terms of equality, democracy and human rights

- **governance**, not only in terms of competences and responsibilities, but also in terms of **power, trust and accountability**;
- **societal, legal and economic** functioning of socio-technical systems;
- **value-based design** approaches and of ethical frameworks;
- **inclusion and diversity** in design, and how such strategies may inform processes and results;
- **distributed and increasingly ubiquitous nature of AI** applications and developing new scholarly perspectives on human-machine communication.