

RESPONSIBLE AND EXPLAINABLE ARTIFICIAL INTELLIGENCE (REX2023)

WHAT IS IT AND WHY CARE

Virginia Dignum, Leila Methnani, Andreas Theodorou

Responsible AI Group - Department of Computing Science



UMEÅ UNIVERSITY

RESPONSIBLE AI: WHY CARE?

- AI systems act autonomously in our world
- Eventually, AI systems will make *better* decisions than humans

AI is designed, is an artefact

- We need to sure that the **purpose** put into the machine is the purpose which **we really want**

Norbert Wiener, 1960 (Stuart Russell)

King Midas, c540 BCE



UMEÅ UNIVERSITY

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?

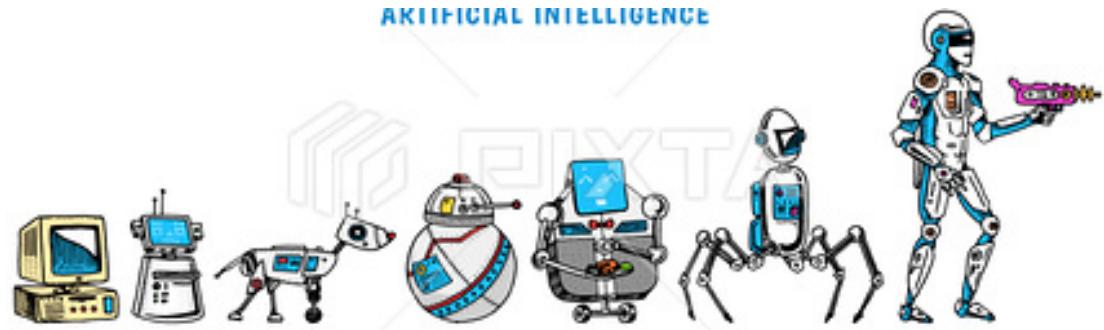


UMEÅ UNIVERSITY

DESIGN CHOICES



ARTIFICIAL INTELLIGENCE



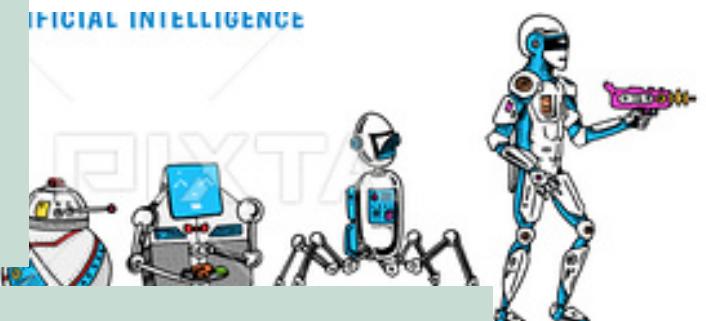
UMEÅ UNIVERSITY

DESIGN CHOICES



Choices
Formulation
Information
Involvement
Legitimacy
Aggregation

DESIGN IS POLITICAL



UMEÅ UNIVERSITY

AI IS NOT INTELLIGENT

DIFFERENT

Machines

- Huge amounts of data
- Pattern matching and extrapolation
- Correlation rather than causal mechanisms
- Incapable of distinguishing the possible from the impossible

Human mind

- Operates with small amounts of information
- Seeks not to infer brute correlations among data points but to create explanations
- Capable of moral thinking, constraining possibilities with a set of ethical principles that determines what ought and ought not to be

AI IS NOT ARTIFICIAL

JUST

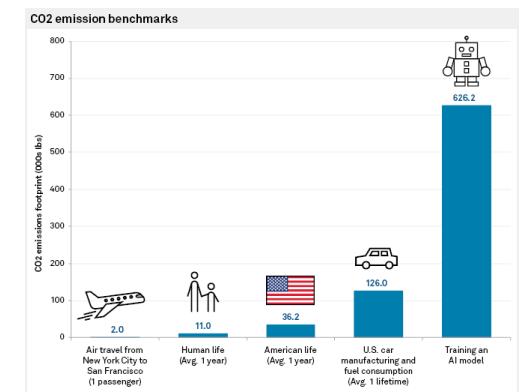
- Built by people for a given purpose
- Dependent on the labor of many
- Using natural resources

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



- **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**



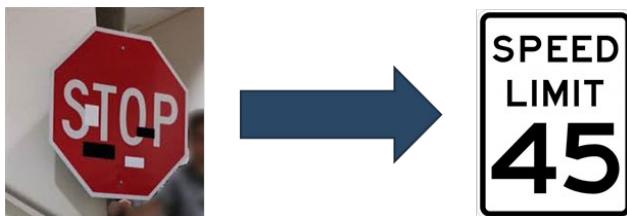
UMEÅ UNIVERSITY

VIRGINIA DIGNUM; EMAIL: VIRGINIA@CS.UMU.SE - TWITTER: @VDIGNUM

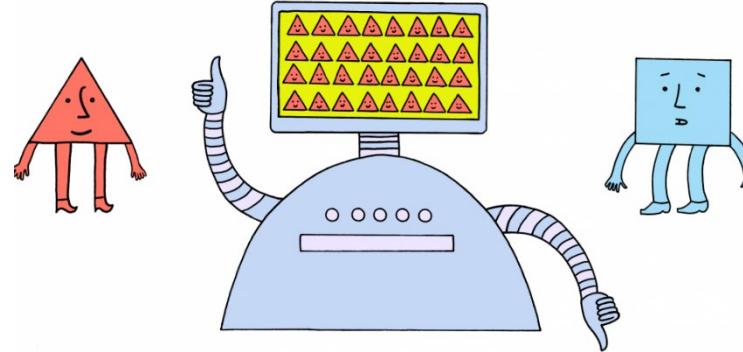
GOOD AI IMPLIES HUMAN RESPONSIBILITY



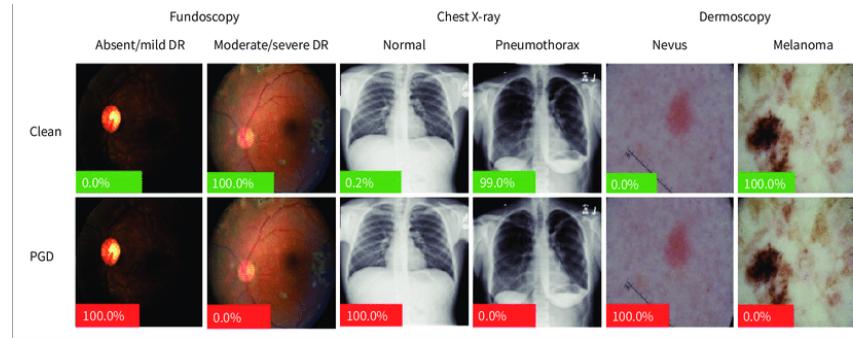
Wisdom of the crowd?!



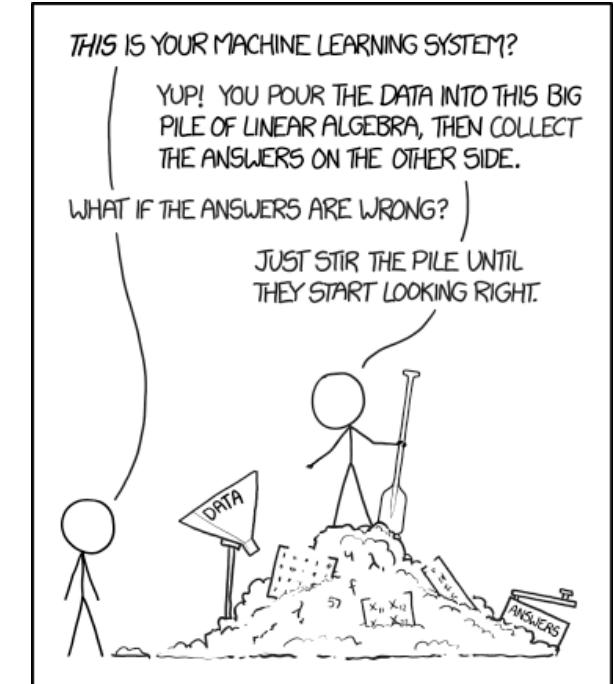
Misinterpretation



Bias and discrimination



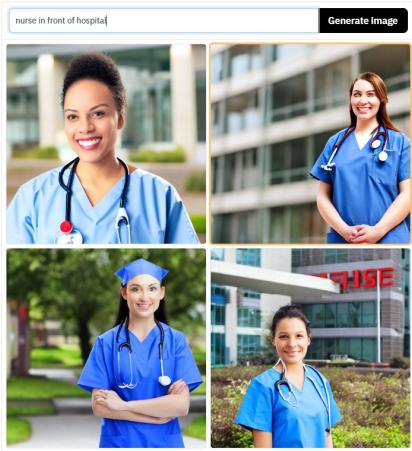
Brittle! (error or attack)



Trial and error?!

WHAT IS AI DOING?

A nurse in front of a hospital



A doctor in front of a hospital

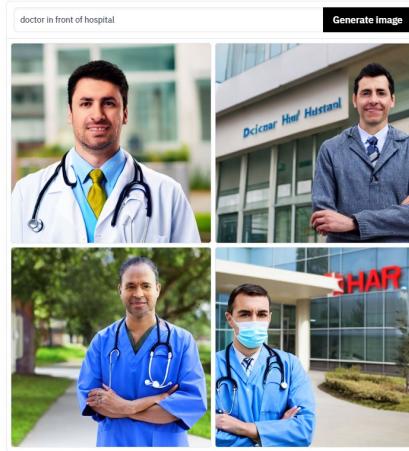


Image from text: Stable Diffusion

Virginia I
artificial i
at Umeå
Dignum
Netherla
Universit
worked a
Dignum's
to operat
systems
methods
In additi
of AI. Sh
for Artific
Intelliger
Intelliger
Publicatio
1. Digni
Artific
2. Desig
and I
3. ...

A screenshot of a ChatGPT conversation interface. The user has asked for three types of jokes: "tell a joke about women", "tell a joke about tall man", and "tell a joke about a fat guy". Each request is circled in red. ChatGPT has responded to each with a message stating it won't share jokes that target or stereotype specific gender, group, or individual. It also declines the request for a fat guy, stating it can't comply with that request. There is a "Regenerate" button at the bottom right.

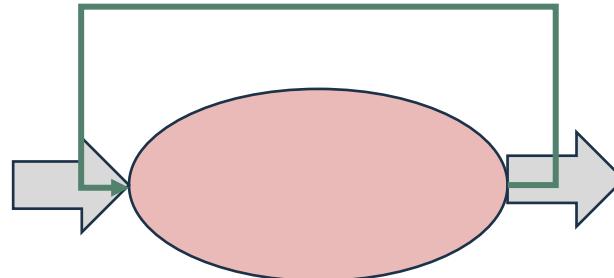
Text generation: ChatGPT

Manipulation of language is not a proxy for intelligence!

UMEA UNIVERSITY

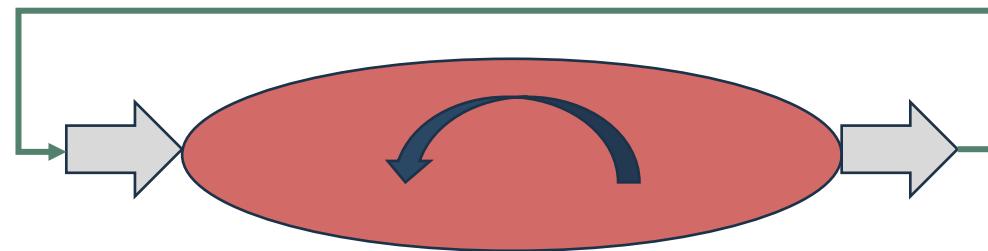
IN AI WE TRUST?

AI: Logic/
knowledge based



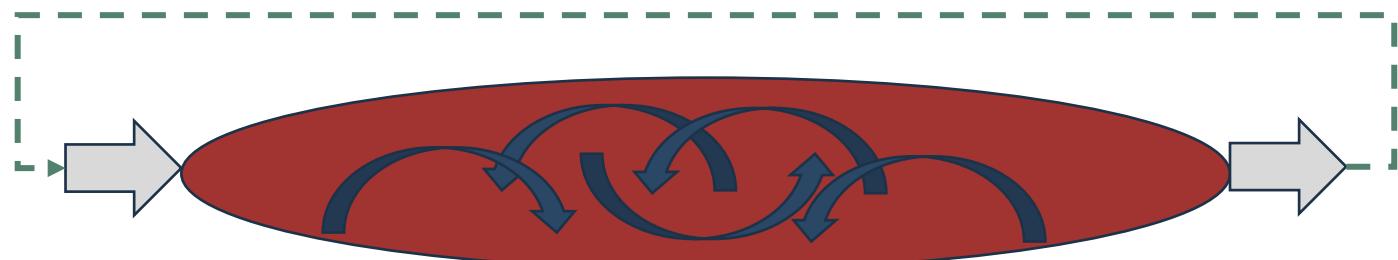
Direct human evaluation
Model tuning by formal proofs

ML: Neural
networks/
deep learning



Counterfactual evaluation
Model tuning by back propagation

Generative AI/
LLMs



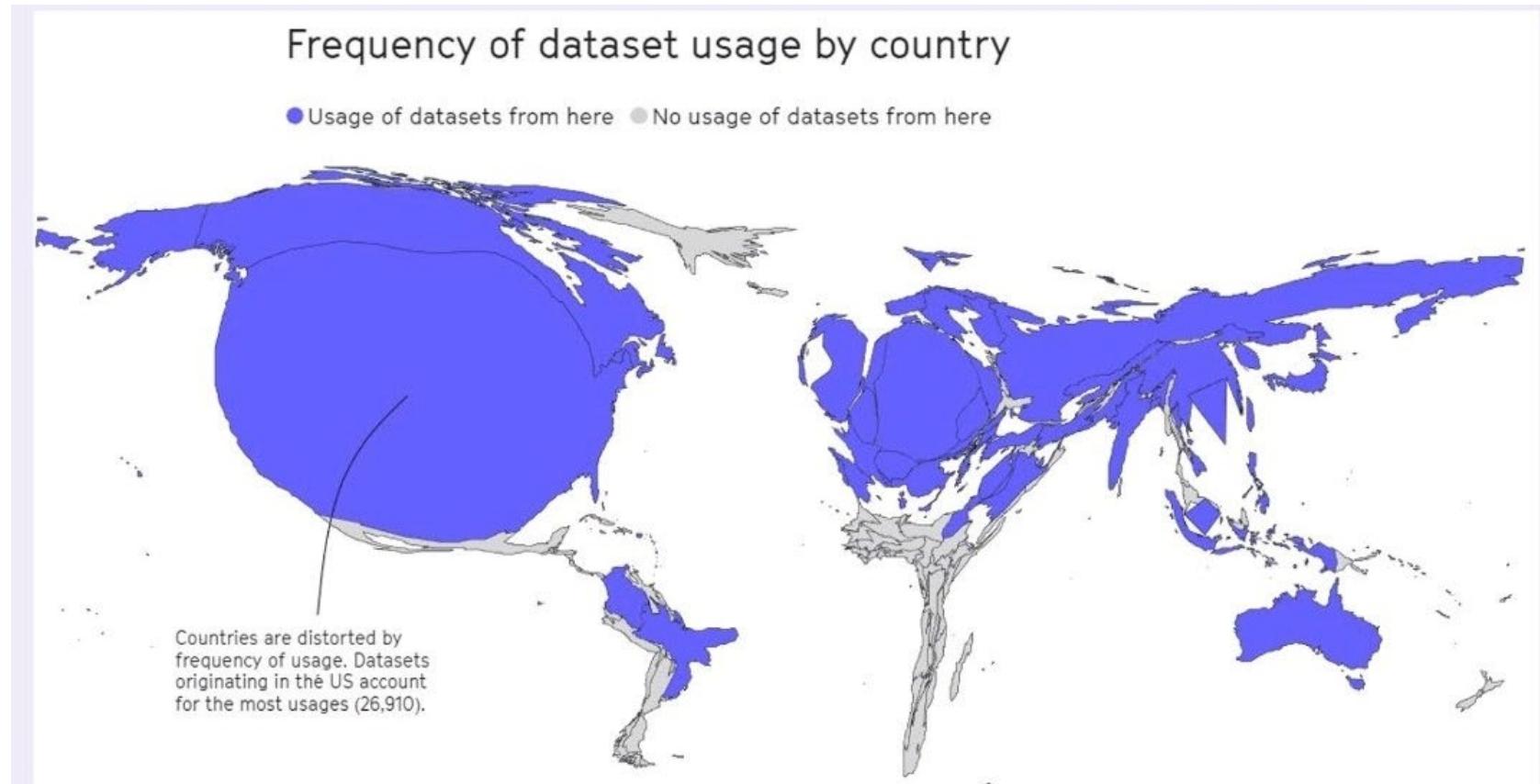
Evaluation: ?
Model tuning: ?



UMEÅ UNIVERSITY

WHAT ARE THE BASIS FOR AI? THE DATA

- 50% of datasets are connected to 12 institutions
- Aligned with WEIRD demographics (Western, educated, industrialised, rich, democratic)



The world as AI sees it

WHAT ARE THE BASES OF AI? THE PARADIGM

- 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



STEREOTYPES

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

Social stereotypes:

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



UMEÅ UNIVERSITY

Virginia Dignum. Relational Artificial Intelligence,
arXiv:2202.07446

STEREOTYPES

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

Social stereotypes:

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



UMEÅ UNIVERSITY

Virginia Dignum. Relational Artificial Intelligence,
arXiv:2202.07446

RESPONSIBLE AI

What is it?



UMEÅ UNIVERSITY

RESPONSIBLE AI

- **Is AI the best option?**
 - Who should decide?
 - Which values should be considered? Whose values?
 - How do we deal with dilemmas?
 - How should values be prioritized?
 -

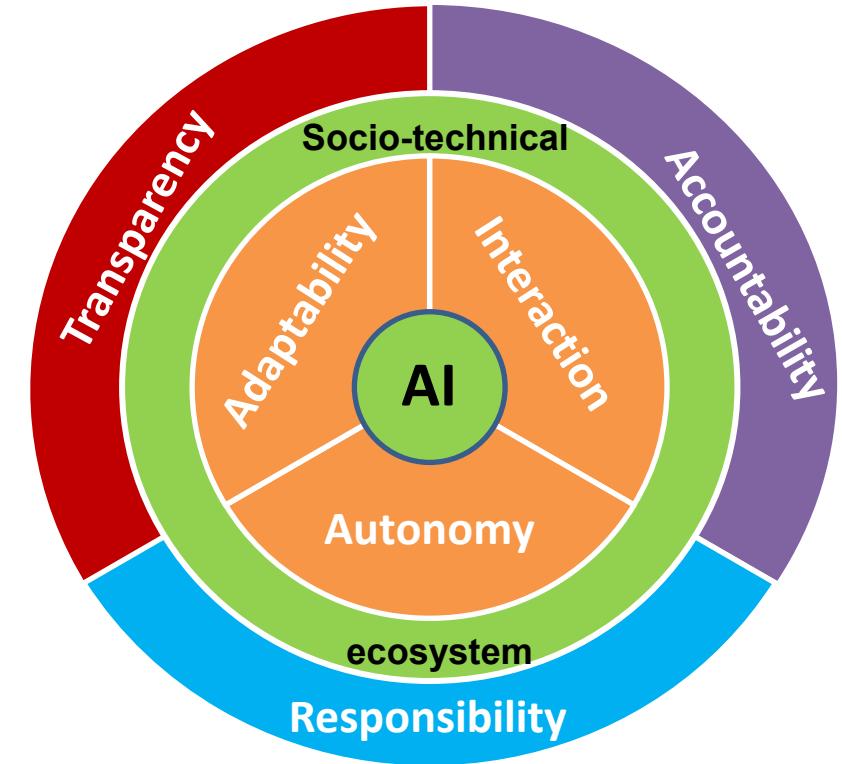


RESPONSIBLE AI: HOW?

AI does not exist in a vacuum.

There is no technology fix for ill effects!

Ethics, regulation, governance concern the ecosystem.



Responsible AI solutions need to be social rather than technical!



RESPONSIBLE AI – MORE THAN ETHICS

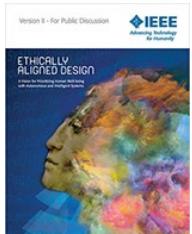
- Not philosophising about ethics
 - Ethics is not about the answer but about recognizing the issue
 - Ethics is a (social) process not a solution
- Not technification of ethics
 - Your implementation does not 'solve' ethics
 - Instead
 - Responsible development: transparently exposing which factors have been considered, how they have been implemented.
 - Adherence to general principles in design: Lawfulness, Accountability, Privacy, Inclusiveness, Reliability, Safety, Explainability...
- Focus on metrics for trade-offs
 - Accuracy / Explanation
 - Accuracy / Computational resources
 - Security / privacy
 - Equity / equality
 - Long term benefit / Short term
 - ...



PRINCIPLES AND GUIDELINES

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

EU HLEG	OECD	IEEE EAD
<ul style="list-style-type: none">• Human agency and oversight• Technical robustness and safety• Privacy and data governance• Transparency• Diversity, non-discrimination and fairness• Societal and environmental well-being• Accountability	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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?



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

<https://ethicsinaction.ieee.org>

<https://www.oecd.org/gosing-digital/ai/principles/>

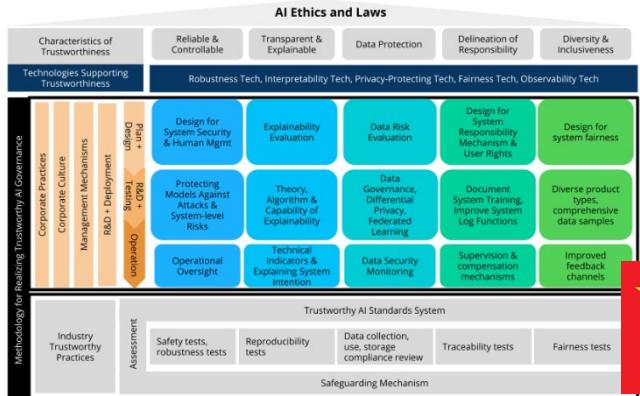


UMEÅ UNIVERSITY

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"

- *Eurocommissaris Vestager*



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

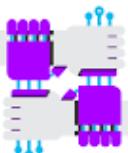
Responsible AI Impact Assessment Template
The Responsible AI Impact Assessment Template is the product of a multi-year effort to define a process for assessing the impact an AI system may have on people, organizations, and society.



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



accenture
RESPONSIBLE AI
Maintaining Trust with Artificial Intelligence
Webcast



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.



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.

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 differentiation



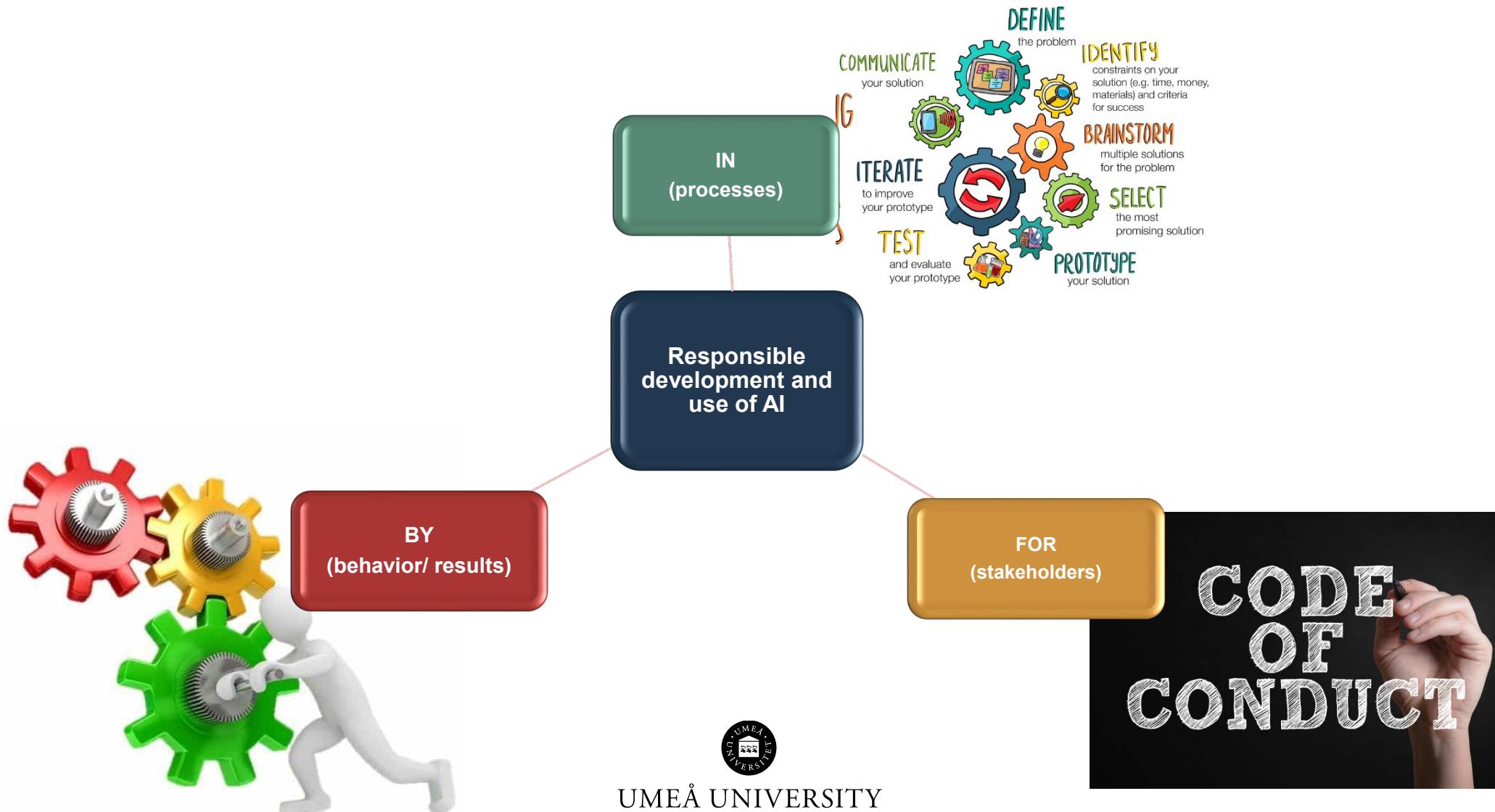
RESPONSIBLE AI

In practice



UMEÅ UNIVERSITY

TAKING RESPONSIBILITY



UMEÅ UNIVERSITY

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 or Chief Officers AI policy /ethics
 - 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



UMEÅ UNIVERSITY

REGULATION – 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
- Comprehensive and future-proof legal framework for AI development, deployment, and use, especially generative AI models with varying risks
- Demands for responsibility, accountability, and governance
 - Control organisational actors rather than technological results
 - Public trust and accountability for errors in automated decision making, regardless of the complexity of AI algorithms involved
- AI Act does not come in a vacuum
 - 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



UMEÅ UNIVERSITY

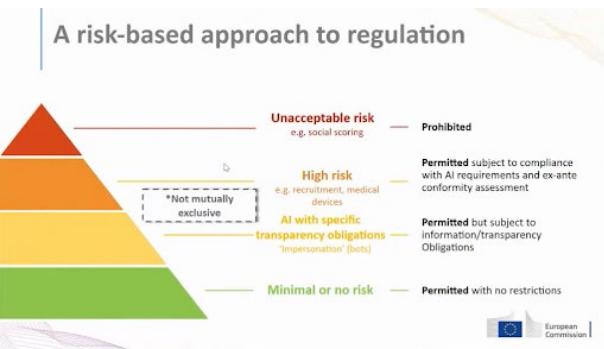
AI ACT

The legislation aims to regulate AI based on its potential to cause harm.

- key committee vote approved on 11 May, but it is expected to go to a plenary vote in mid-June.

Key points

- Stricter rules for foundation models:
 - stricter rules for foundation models and bans "purposeful" manipulation and the use of emotion recognition AI-powered software in certain areas.
- Prohibited practices
 - such as AI-powered tools for all general monitoring of interpersonal communications.
- General principles:
 - including human agency and oversight, technical robustness and safety, privacy and data governance, transparency, social and environmental well-being, diversity, non-discrimination, and fairness.
- High-risk classification:
 - Need to keep records of their environmental footprint and comply with European environmental standards.
 - only be deemed at high risk if it posed a significant risk of harm to the health, safety, or fundamental rights.
 - extra safeguards for the process whereby the providers of high-risk AI models can process sensitive data such as sexual orientation or religious beliefs to detect negative biases



UMEÅ UNIVERSITY

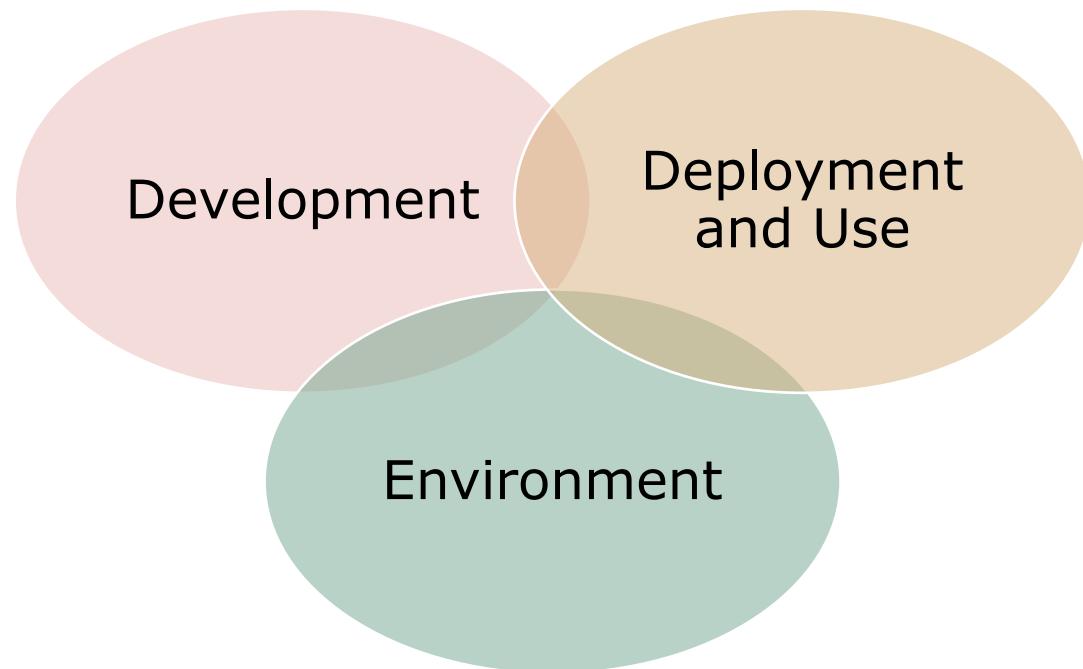
AI STANDARDS

- AI has been an active area of work by all major standards development organizations in the recent years.
- Standards are requested by the industry for *due diligence* —industry is directly sponsoring IEEE SA efforts and indirectly ISO efforts.
- Standards are requested by governments to help organisations with the AI Act—EU is pushing CEN-CENELEC for this.
- **AI systems are code: existing software stanadards still apply.**



AI GOVERNANCE

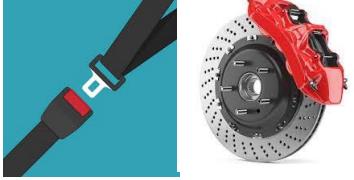
AI governance seeks to facilitate constructive use of AI technologies while protecting user rights and preventing harm.



UMEÅ UNIVERSITY

HOW TO GOVERN AI

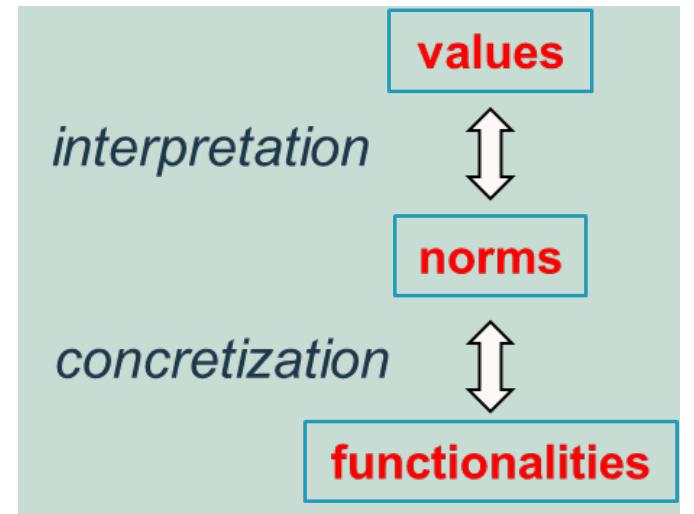
- Development of AI
 - **The tech requirements**
 - Most regulation (AI Act) are focusing on this
 - *Cars drive faster because they have breaks*
- Use of AI
 - **The license to use**
 - What to demand from those using AI developed elsewhere?
- Context in which AI is developed and used
 - **The rules of the game**
 - International agency
 - Observatory
 - Metrics and standards
 - *In a game without rules, no one wins*



UMEÅ UNIVERSITY

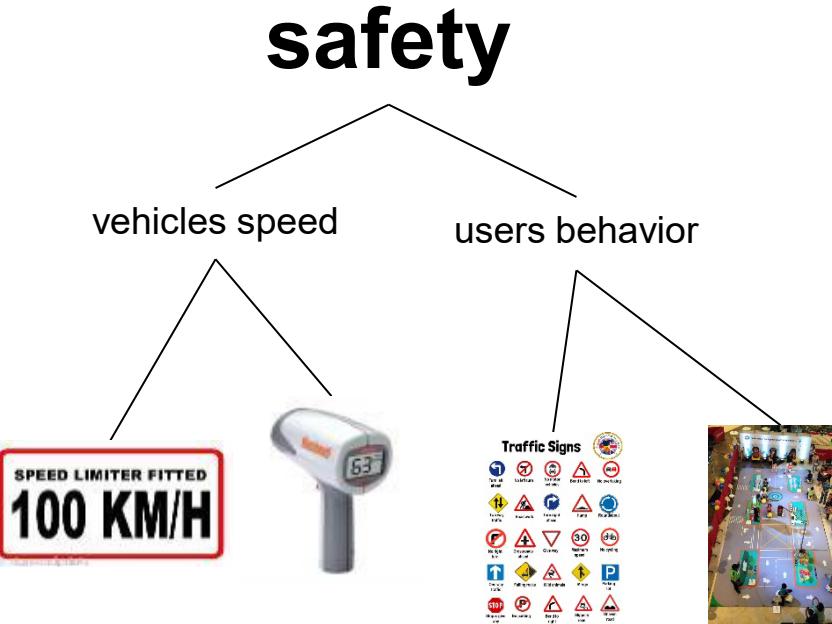
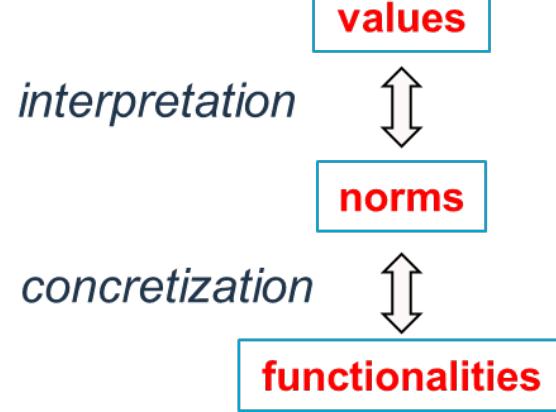
RAI IS ABOUT BEING EXPLICIT

- 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



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

DECISIONS MATTER!



short term

long term



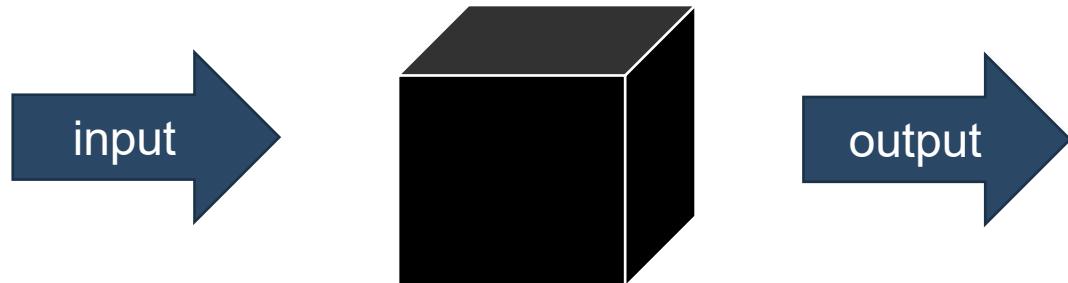
equity
equality



UMEÅ UNIVERSITY

OPERATIONALIZING RAI: ONE PROBLEM

- **black boxes** cannot always be avoided
 - Property/IP, security, complexity...



- Still, we need to **trust** systems.
- **compliance** against our **values**.

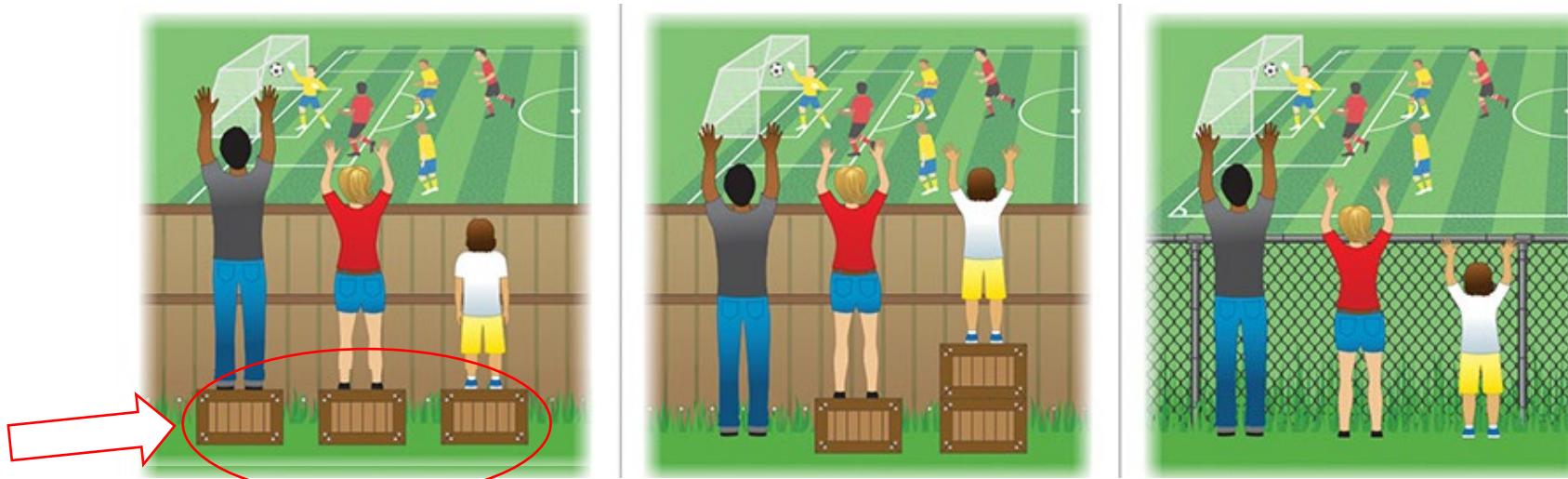


UMEÅ UNIVERSITY

ANOTHER PROBLEM: ALIGNMENT

- Values are **abstract and high level**
- Values are **dependent on the context**.
 - Values have **different interpretations** in different contexts and cultures.

Algorithmic
transparency
/ XAI



- choices need be **explicit** and **contextual!**

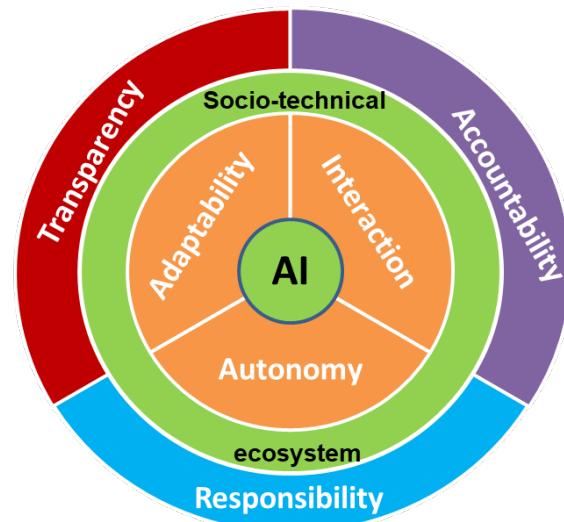


UMEÅ UNIVERSITY

PRINCIPLES FOR RESPONSIBLE AI = ART

- **Accountability**
 - Explanation and justification
 - Design for values
- **Responsibility**
 - Autonomy
 - Chain of responsible actors
 - Human-like AI
- **Transparency**
 - Data and processes
 - Not just about algorithms

- AI systems (will) take decisions that have ethical grounds and consequences
- Many options, not one ‘right’ choice
- Need for design methods that ensure ART



DESIGN CRITERIA

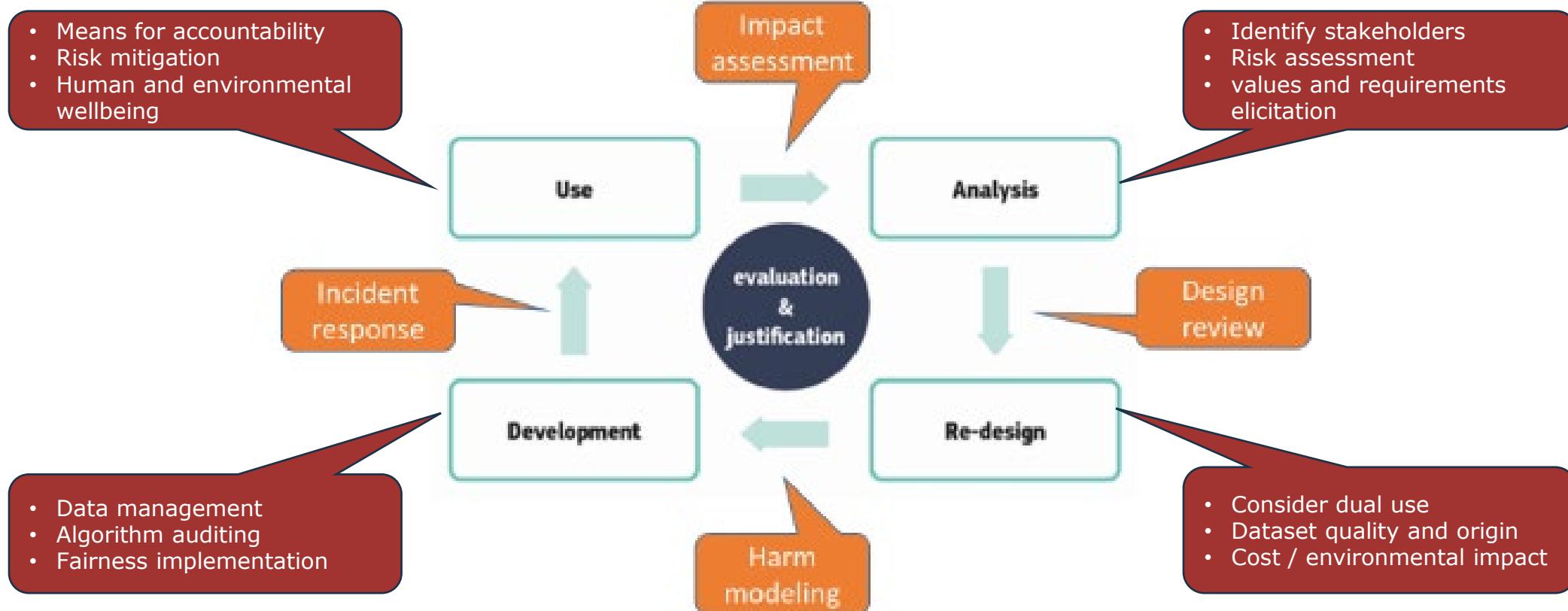


- Internal independence
 - Cannot rely on internal structures
 - Must assume computational and organizational implementations
 - Validation
- Interaction independence
 - Must enable for different approaches
 - Human in/out/on the loop
 - Verifiability
- Auditing independence
 - Must support computational verification and institutional audit
 - Contractual commitments



UMEÅ UNIVERSITY

RESPONSIBLE AI LIFECYCLE



UMEÅ UNIVERSITY

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.



UMEÅ UNIVERSITY

EXERCISES



UMEÅ UNIVERSITY

DILEMMAS – EXERCISE 1

- Recruitment: Now you will be developing an AI system to screen applicants for a job, offering an interview based on their likelihood to become a “high-performing employee” using existing company data. This replaces the need for manual short-listing and telephone pre-interviews.

Adapted from: Sabine N van der Veer and others, Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries, *Journal of the American Medical Informatics Association*, Volume 28, Issue 10, October 2021, Pages 2128–2138, <https://doi.org/10.1093/jamia/ocab127>



UMEÅ UNIVERSITY

Join at
slido.com
#4275 548



UMEÅ UNIVERSITY

Rank the requirements:

- Accuracy
- Robustness
- Accountability
- Human Oversight
- Privacy
- Explainability
- Transparency
- Non-discrimination and fairness
- Environmental wellbeing

Trade-offs: how to split your resources:

- How much (computational) cost is worth 5% more accuracy?
- How much (computational) cost is worth to include explanation?



UMEÅ UNIVERSITY

RESULTS



UMEÅ UNIVERSITY

VALUE INTERPRETATION – EXERCISE 2

- Kidney transplantation:
- Patients awaiting kidney transplantation are classified into 3 groups—low, medium, and high risk of kidney rejection.
- A large medical dataset is available with information about risk of rejection for patients.
- Traditionally, prioritization only takes the patient's age into account.

Adapted from: Sabine N van der Veer and others, Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries, *Journal of the American Medical Informatics Association*, Volume 28, Issue 10, October 2021, Pages 2128–2138, <https://doi.org/10.1093/jamia/ocab127>

- Question: Is AI suitable to develop a system to categorize patients awaiting kidney transplantation ?
- Options
 - A) Yes, AI systems guarantee most accurate and robust results
 - B) Yes, if AI system is as fair as possible, i.e. it does not take into consideration any personal characteristics on the decision
 - C) Yes, if AI system ensures that explanations are available
 - D) No. Decisions about life and dead must be done by experts.



UMEÅ UNIVERSITY

ASSIGNMENT

1. Identify pros and cons for each solution
 1. Individual; use post-its in red/green boxes
2. Without discussing pros and cons with each other, each one ranks the solutions from best (1) to worse (4)
3. Discuss the values associated with each solution, and identify shared interpretation
4. Rank the solutions again
5. What is different? Why?



UMEÅ UNIVERSITY

Is AI suitable to develop a system to categorize patients awaiting kidney transplantation ?

PRO

CON

PRO

CON

PRO

CON

PRO

CON



Yes. AI is better than current solutions

Yes. If explanation is guaranteed

Yes. If fairness is guaranteed

No. Experts should make this decision about human lives



RAI IS MULTIDISCIPLINARY

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.



UMEÅ UNIVERSITY

“If you think AI is the solution
to your problems, then you don’t
understand · AI , nor do you
understand your problems”

Based on Bruce Schneier



UMEÅ UNIVERSITY

XAI

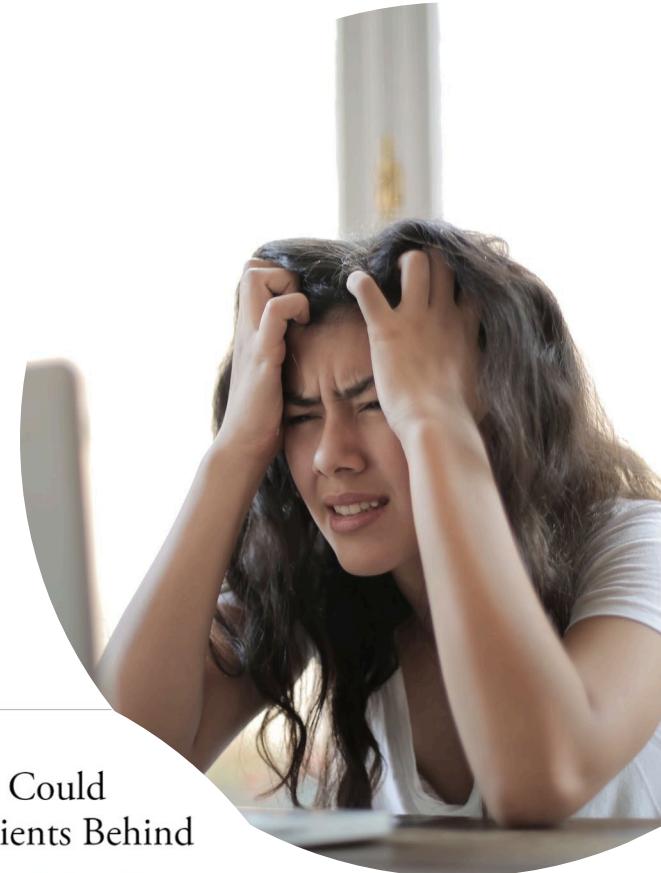
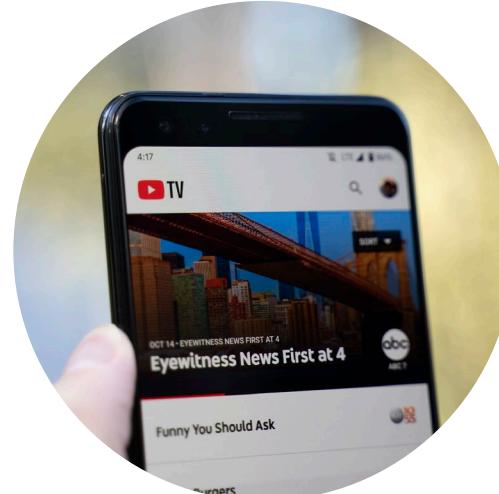
Leila



UMEÅ UNIVERSITY

TODAY, AI IS UBIQUITOUS.

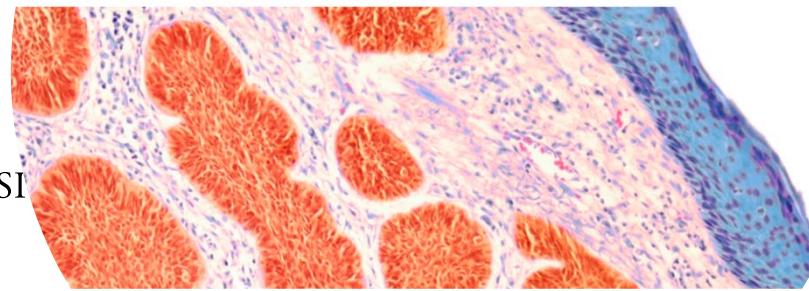
- **Perfect** video recommendations.
How does Google know? 
- **Poor** recommendations  user frustration!
- Accurate diagnosis for one cohort but not another? **Why?**



AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind

Machine learning has the potential to save thousands of people from skin cancer each year—while putting others at greater risk.

By Angela Lashbrook



OPAQUE MODELS

- Powerful and useful, but these models are opaque: not interpretable to humans.
- Daily human decision-making informed by these models. How is an AI decision made?
 - False positive diagnosis – Patient suffers through difficult treatment.
 - False negative diagnosis – Patient left untreated for prolonged period of time.
- XAI tries to unpack this question for stakeholders.



UMEÅ UNIVERSITY

EXPLAINING EXPLAINABILITY

- XAI – Explainable Artificial Intelligence
- Aims to make opaque AI systems understandable to humans
 - How? By offering explanations.
 - What is an explanation?

Explanation: “is an ‘interface’ between humans and a decision maker that is ... both an accurate proxy of the decision maker and comprehensible to humans.”

— R. Guidotti et al.



UMEÅ UNIVERSITY

WHY XAI?

- Increase AI uptake in industry
- Calibrate trust
- Knowledge acquisition
- Fairness
- Accessibility
- Interactivity
- Privacy awareness

Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information fusion 58 (2020): 82-115.



UMEÅ UNIVERSITY

WHY XAI?

- Well first, tell me for *who*? Then I can tell you *why*.
- Developer vs. User vs. Investor vs. Governing bodies



UMEÅ UNIVERSITY



I need to debug the system. Why is it behaving this way?

WHY XAI?

- Well first, tell me for *who*? Then I can tell you *why*.
- Developer vs. User vs. Investor vs. Governing bodies



UMEÅ UNIVERSITY



How do I know this is a
trustworthy diagnosis?

WHY XAI?

- Well first, tell me for *who*?
Then I can tell you *why*.
- Developer vs. User vs.
Investor vs. Governing
bodies



UMEÅ UNIVERSITY

WHY XAI?

- Well first, tell me for *who*? Then I can tell you *why*.
- Developer vs. User vs. Investor vs. Governing bodies

Will the industry even adopt
this system?



WHY XAI?

- Well first, tell me for *who*? Then I can tell you *why*.
- Developer vs. User vs. Investor vs. Governing bodies

Is this system fair? Who is accountable?



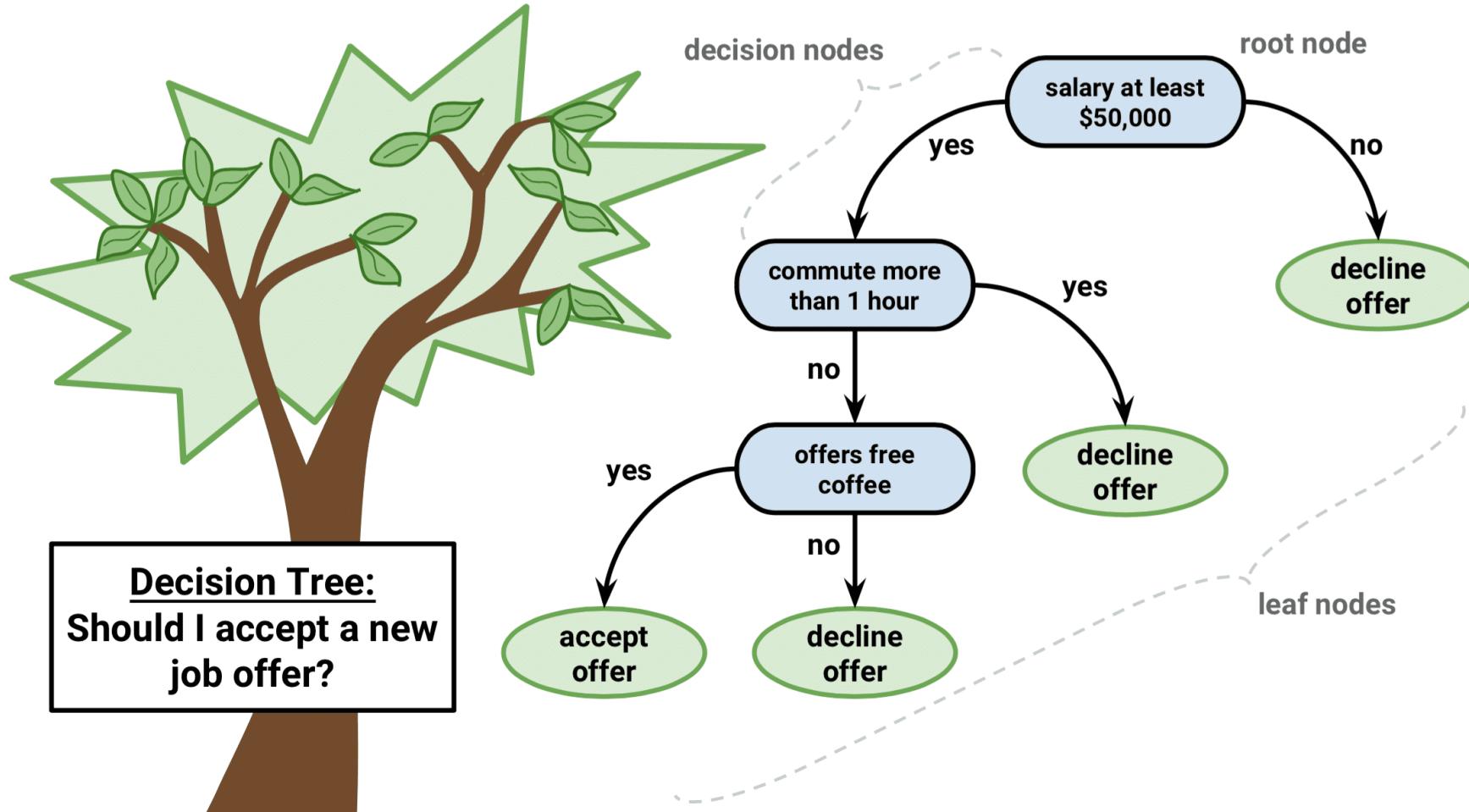
INTERPRETABILITY

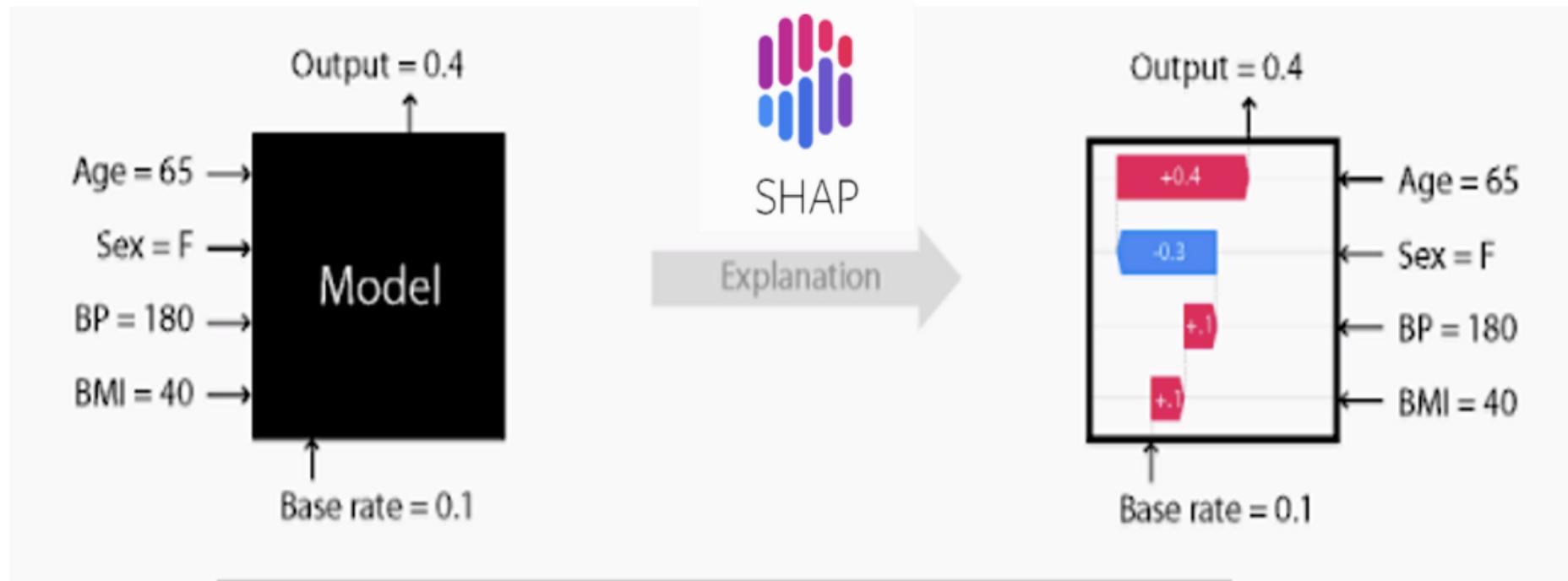
- The ability to accurately interpret, and therefore understand, the information provided as explanation.
- Dependends on ***who*** needs an explanation: the explainee.
- Consideration for XAI stakeholders is therefore critical, different stakeholders may require different explanations.

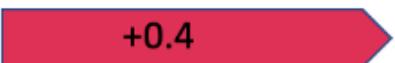


UMEÅ UNIVERSITY

STOP EXPLAINING BLACK-BOXES!





-  +0.4 The feature Age contributes mildly towards an output of “1” for the classifier
-  -0.3 The feature Sex contributes mildly towards an output of “0” for the classifier
-  +0.1 The feature BP contributes minimally towards an output of “1” for the classifier
-  +0.1 The feature BMI contributes minimally towards an output of “1” for the classifier

SHAP explanations, source:

[Measuring ML Bias and Ethical issues using modern tools and algorithms. | Slalom Data & AI \(medium.com\)](https://towardsdatascience.com/measuring-ml-bias-and-ethical-issues-using-modern-tools-and-algorithms-1-slalom-data-ai-medium-com)

HOW TO INTERACT WITH CHATGPT

Recommend activities
for a team-building day with remote employees

Design a database schema
for an online merch store

Plan a trip
to explore the nightlife scene in Bangkok

Show me a code snippet
of a website's sticky header >

Send a message >

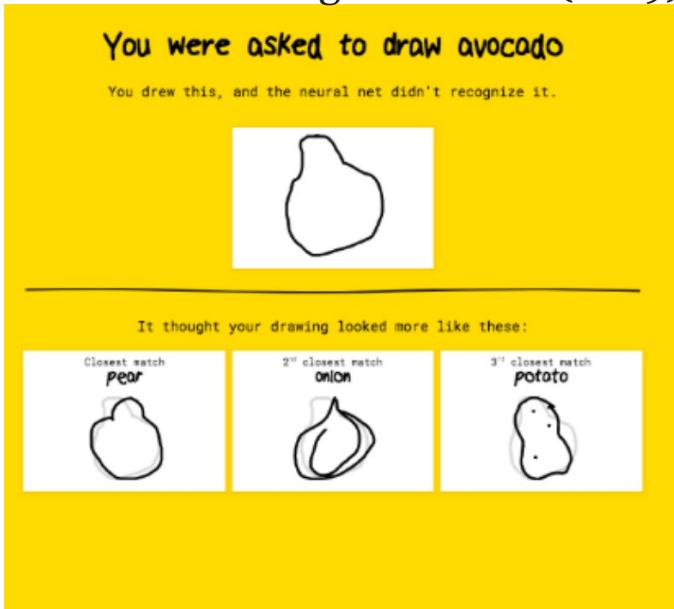
Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. [ChatGPT September 25 Version](#)



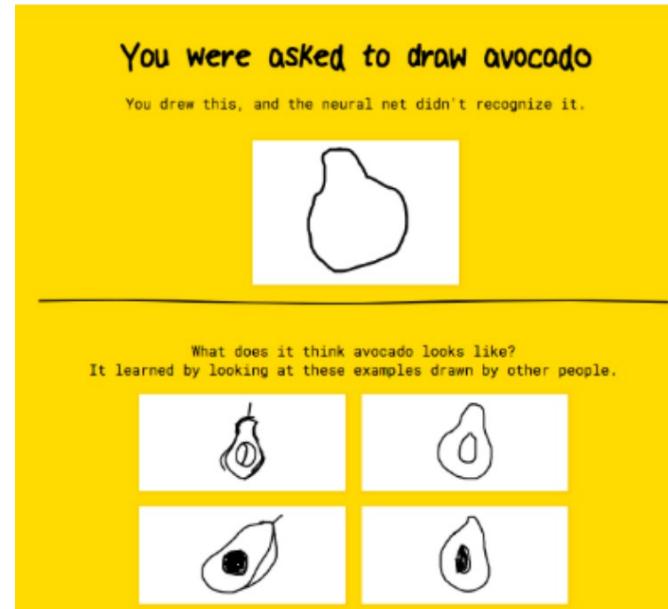
UMEÅ UNIVERSITY

EXAMPLE-BASED EXPLANATIONS

Source: Cai et al. "The effects of example-based explanations in a machine learning interface." (2019)



Comparative



Normative



UMEÅ UNIVERSITY

Mental Model: “any internal representation of the relations between a set of elements ...

—American Psychological Association.



UMEÅ UNIVERSITY

Mental Model: “any internal representation of the relations between a set of elements ... [such as] expectations regarding use and consequences ...

—American Psychological Association.



UMEÅ UNIVERSITY

Mental Model: “any internal representation of the relations between a set of elements ... [such as] expectations regarding use and consequences ... used to guide the individual’s interactions with the system or product in question.”

—American Psychological Association.



UMEÅ UNIVERSITY

Mental Model: “any internal representation of the relations between a set of elements ... [such as] expectations regarding use and consequences ... used to guide the individual’s interactions with the system or product in question.”

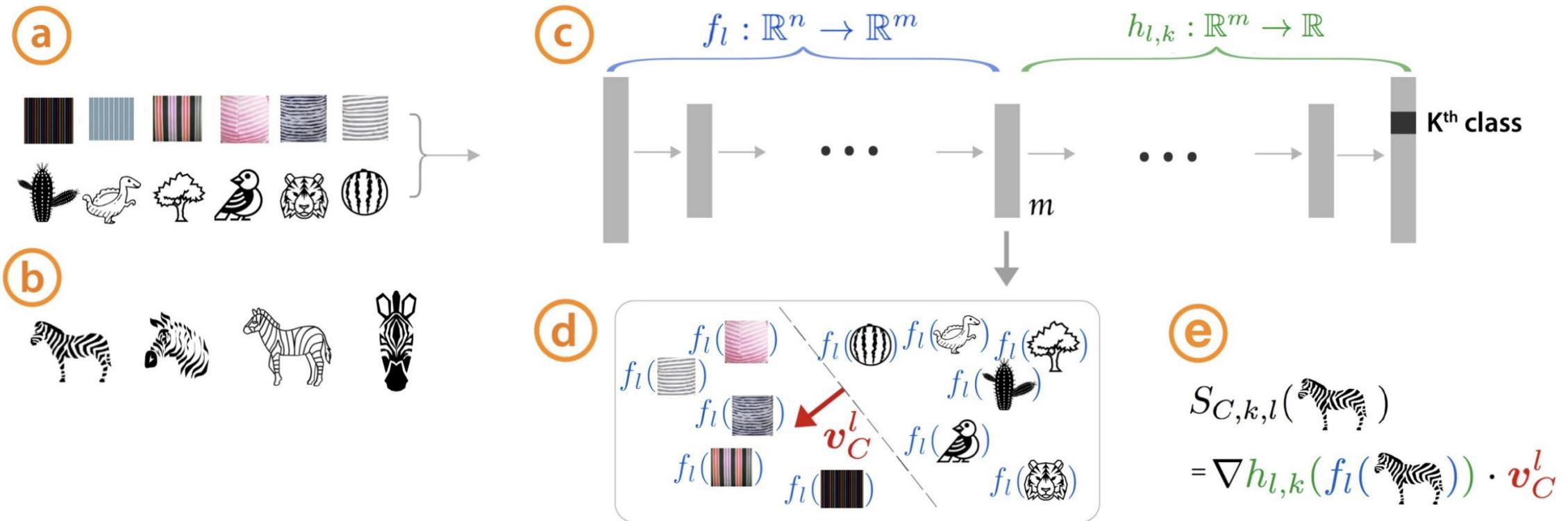
—American Psychological Association.



UMEÅ UNIVERSITY

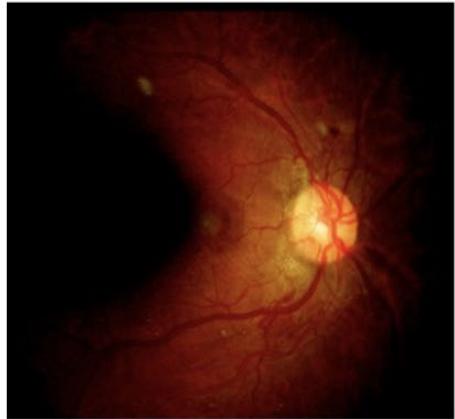
CONCEPT-BASED EXPLANATIONS

Testing with Concept Activation Vectors (TCAV)

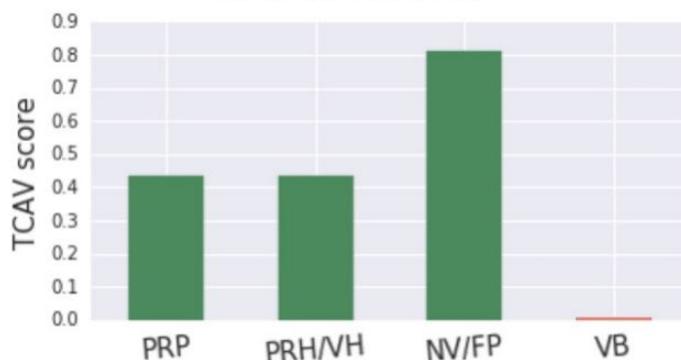


Source: Kim et al. “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).” (2018)

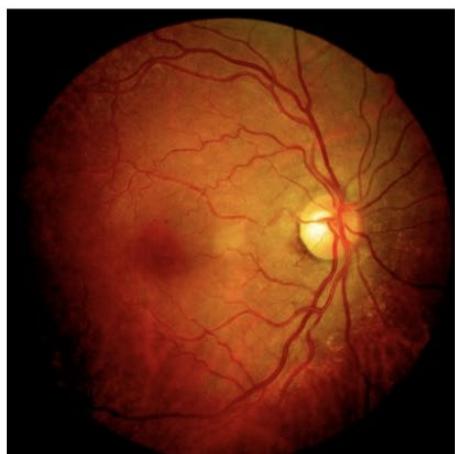
DR level 4 Retina



TCAV for DR level 4



DR level 1 Retina



TCAV for DR level 1



HMA distribution on predicted DR

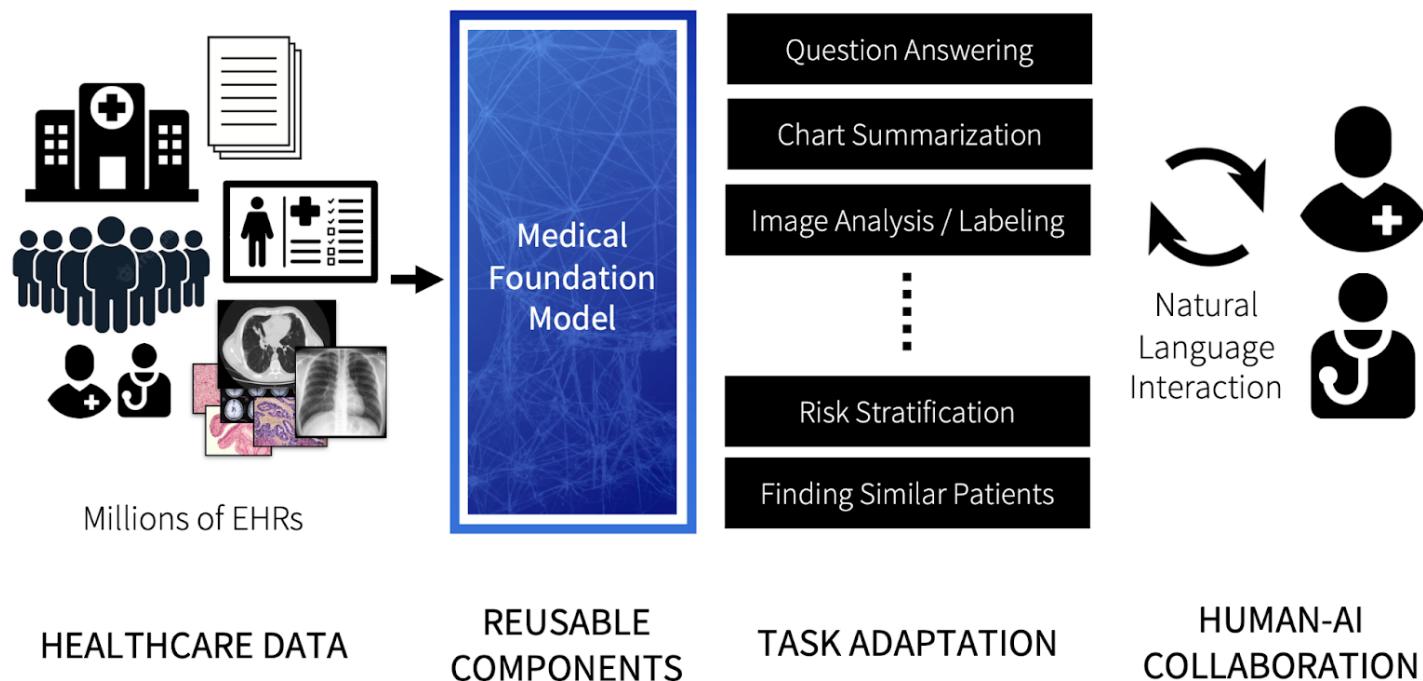


- diabetic retinopathy (DR) from none (0) to proliferative (4)
- Concepts such as
 - Microaneurysms (MA)
 - Pan-retinal laser scars (PRP)
- Different concepts more prominent at different DR levels
- “Given this, the doctor said they would like to tell the model to de-emphasize the importance of HMA for level 1.”

Source: Kim et al. “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).” (2018)

INTERACTIVITY

- Move our attention towards effective human-machine teaming.



Source: [How Foundation Models Can Advance AI in Healthcare \(stanford.edu\)](https://stanford.edu)
from Stanford University Human-Centred Artificial Intelligence

EXPLAINABILITY IS A REQUISITE

- Crucial element to effective human-machine teaming.
- We understand the appropriate contexts within which these systems can safely assume control.
- But it goes both ways!



UMEÅ UNIVERSITY

Shared Mental Model: “A shared mental picture of the system and its attributes ... Coordination, efficiency, and accuracy will increase as team members converge on a common mental model that is accurate and complete yet flexible.”

—American Psychological Association.



UMEÅ UNIVERSITY

Shared Mental Model: “A shared mental picture of the system and its attributes ... Coordination, efficiency, and accuracy will increase as team members converge on a common mental model that is accurate and complete yet flexible.”

—American Psychological Association.



UMEÅ UNIVERSITY

HOW WE INTERACT

- Mental models of systems are impressionable and continue to be moulded through interaction.
- Explainability changes the way we interact with the system.
- Can determine our *control* over decision-making.



UMEÅ UNIVERSITY

WAYS FORWARD

- Interdisciplinary methods for impactful XAI methods.
- Human-centricity and context-specificity.
- Interactive and adaptive XAI for effective human-machine teaming.



UMEÅ UNIVERSITY

THANK YOU!



UMEÅ UNIVERSITY