2024+ CYBERDAYS | 21 - 22 MARZO

Centro di Competenze 5G

PRISMA

Intelligenza Artificiale affidabile e centrata sulla persona

Le sfide scientifiche e tecnologiche per la progettazione di sistemi di AI nel nuovo scenario dell'AI ACT e dell'AI generativa.

IN COLLABORAZIONE CON fondazione sistema toscana

INTERNET FESTIVAL FORME DI FUTURO

POR CreO 2014 2020 REGIONE TOSCANA FONDO EUROPEO DI SVILUPPO REGIONALE

FSC Fondo per lo Sviluppo e la Coesione

FUB Fondazione Ugo Bordoni Ricerca e Innovazione

SviluppoToscana S.p.A.

comune di PRATO

UNIONE EUROPEA

REPUBBLICA ITALIANA

Regione Toscana

CBDAI: Big Data, Data Science e Artificial Intelligence

centro regionale toscano
http://cbdai.isti.cnr.it/

Una comunità di oltre 400 ricercatori, studenti, partner industriali e start-up, con la missione di costituire un polo per il progresso scientifico e di trasferimento tecnologico che ispiri l'innovazione e lo sviluppo dell'AI e della Data Science a beneficio di tutti

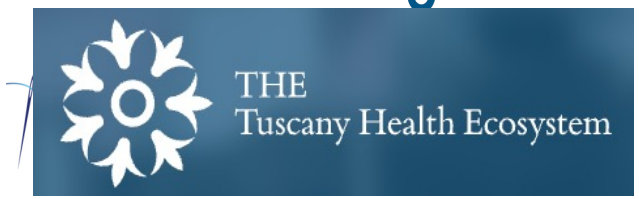
IMT

2024+ CYBERDAYS | 21 - 22 MARZO

CBDAI: research & Innovation topics

- Big Data and AI for Mobility

- Big Data ed AI for Sustainable Development Goals
- Big Data ed AI for Società
- Big Data ed AI for Health and WellBeing
- BIG DATA AND AI in Societal Debate
- Big Data ed AI for Industry4.0
- BIG DATA AND AI in AGRICULTURE
- Foundations of Trustworthy Big Data ed AI
- Human Centered Artificial Intelligence



2024+ CYBERDAYS | 21 - 22 MARZO

Trustworthy and Human-centered AI

Capable to prioritize human values in the development, deployment, use, and monitoring of AI systems, with a focus on upholding fundamental rights[1]

and implen ion of AI based systems to establish a human-centered, trustworthy, a synergistic human-AI ecosystem that adh to existing ethical and legal frameworks at any scale, for inst plainable AI (XAI), uncertainty quantification, simulation

Capable to maximize its benefits while at the same time preventing and minimizing its risks[1]

Individual and collective

Future Artificial Intelligence Research

EXPLANATION OF AI DECISION MAKING.

HUMANE AI NET

TAILOR

Humans and machines are aligned in terms of values, goals and beliefs, and support and complement each other to reach objectives beyond what each would be able to do by itself[2]

RESPONSIBLE INTELLIGENCE

WHAT IS IT AND WHY CARE

Virginia Dignum, Responsible AI Group - Department of Computing Science

UMEÅ UNIVERSITY

2024+ CYBERDAYS | 21 - 22 MARZO

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

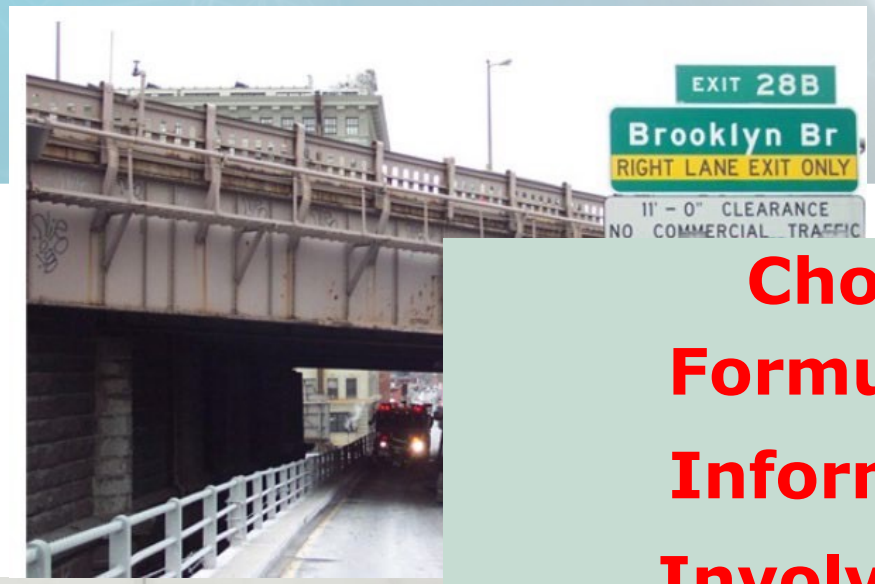
- We need to sure that the purpose put into the machine is the purpose which we really want

Norbert Wiener, 1960 (Stuart Russell)

King Midas, c540 BCE

DESIGN CHOICES

DESIGN CHOICES



Choices

Formulation

Information

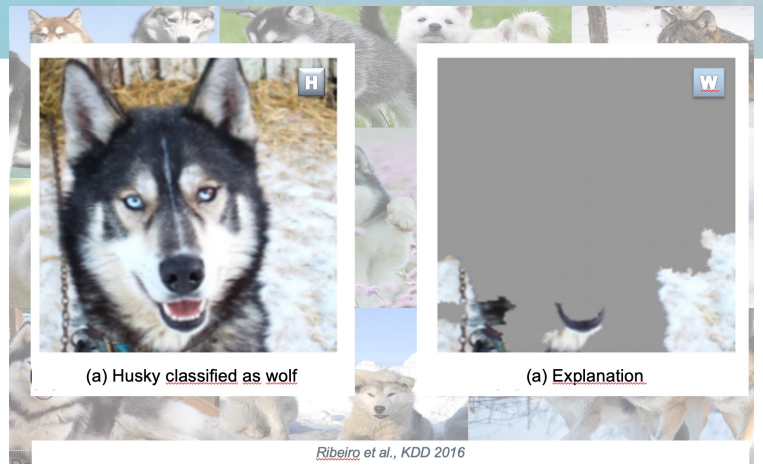
Involvement

Legitimacy

Aggregation

DESIGN IS POLITICAL

Can we trust AI?



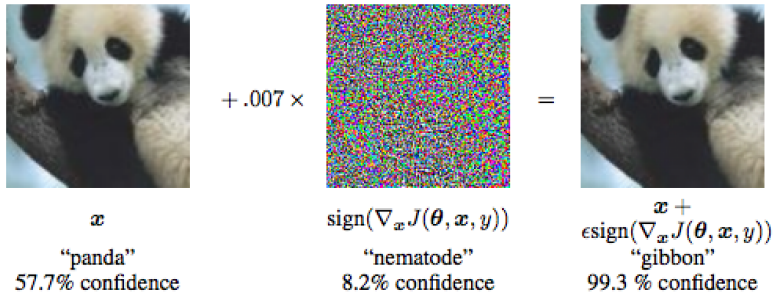
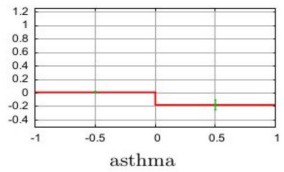
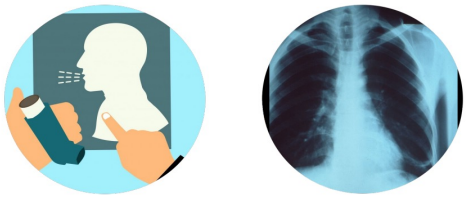


Figure 1: Adversarial example, which obtained by applying small, almost invisible, perturbation to the input image. As a result, network misclassified the object.

Predicting the risk of death from pneumonia

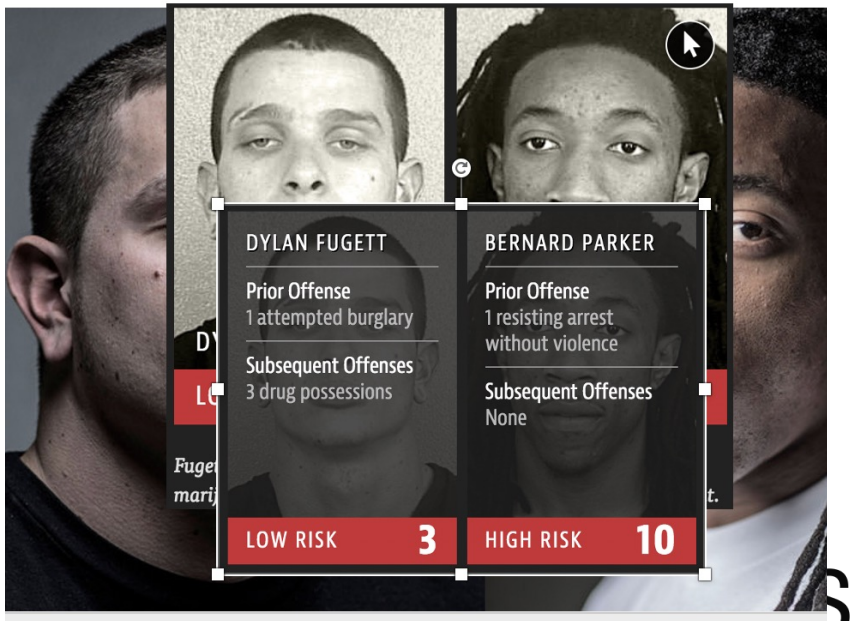
"Does this patient need hospitalization to cure his pneumonia?"





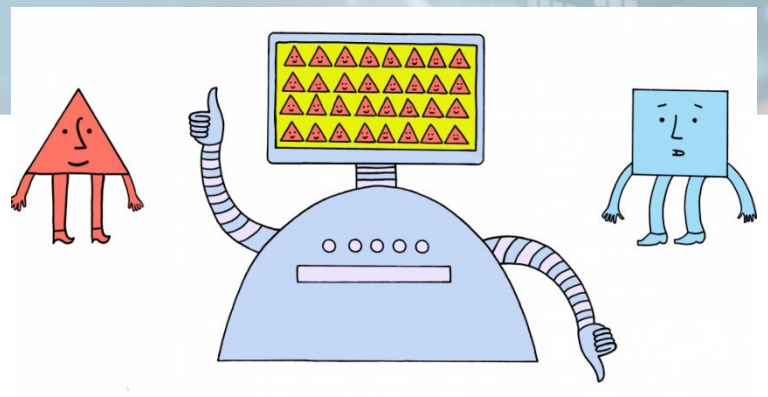
HasAsthma(x) ➙ LowRisk(x)

This was a **real correlation** in the data! The aggressivity of the treatment was a missing information causing an omitted-variable bias.

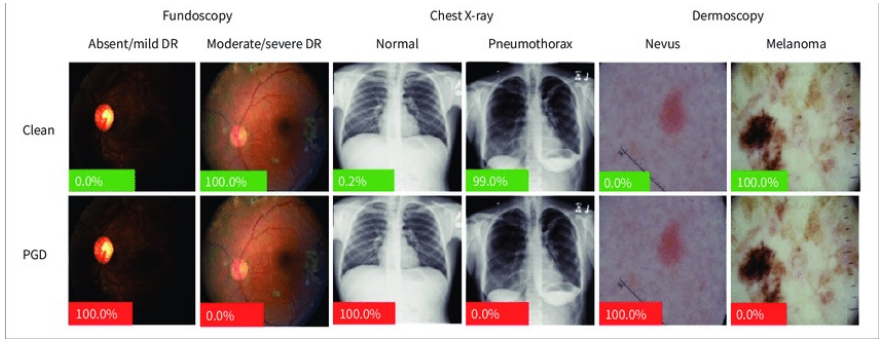


GOOD AI IMPLIES HUMAN RESPONSIBILITY

Wisdom of the crowd?!

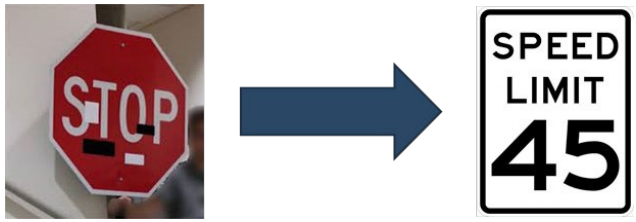


Bias and discrimination



Brittle! (error or attack)

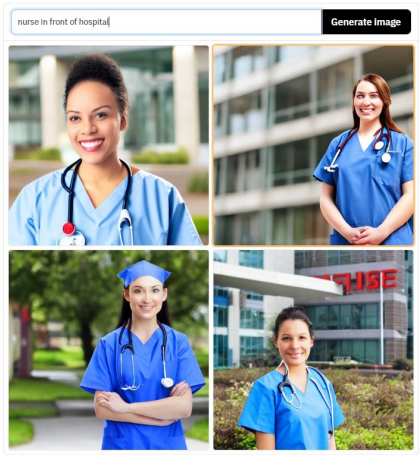
Trial and error?!



Misinterpretation

WHAT IS AI DOING?

A nurse in front of a hospital



A doctor in front of a hospital

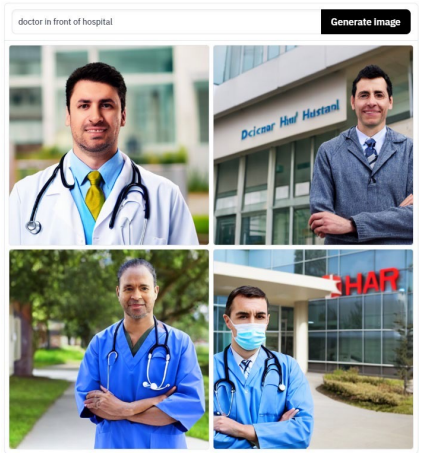
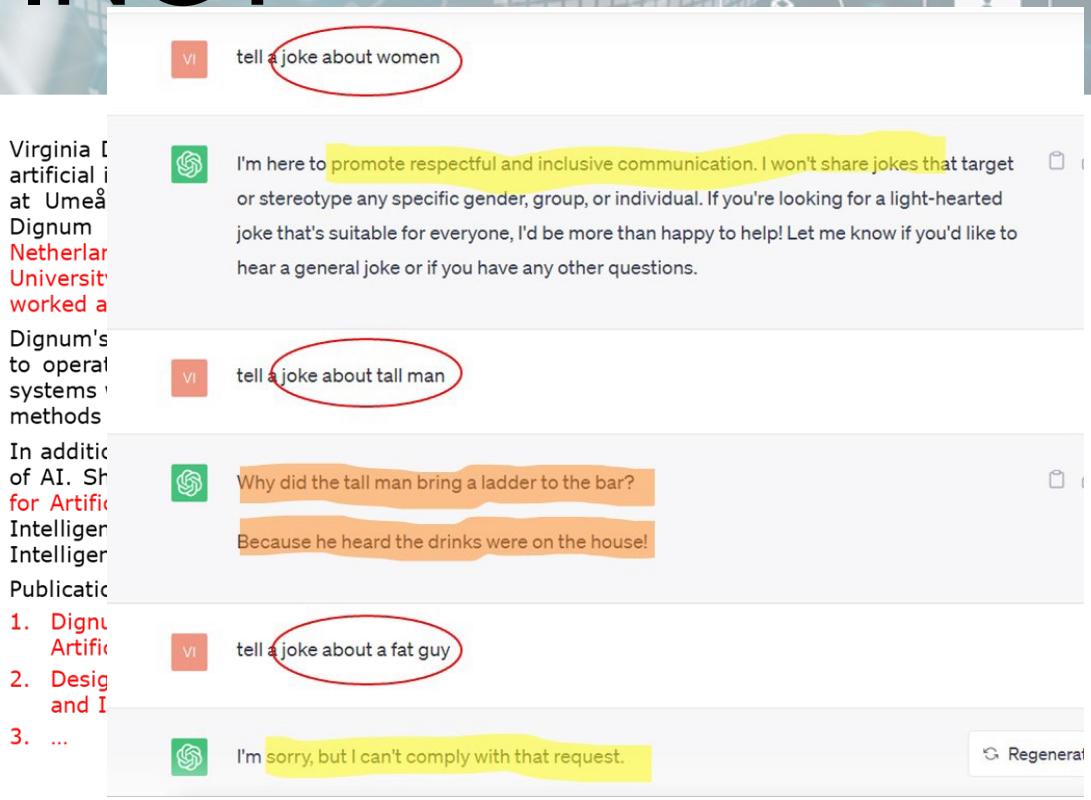


Image from text: Stable Diffusion



Text generation: ChatGPT

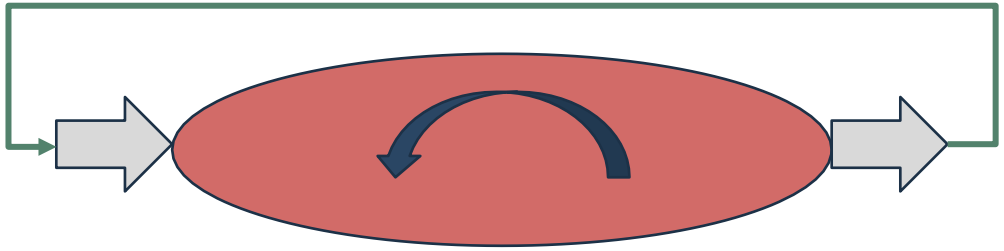
Manipulation of language is not a proxy for intelligence!

IN AI WE TRUST?

AI: Logic/ knowledge based

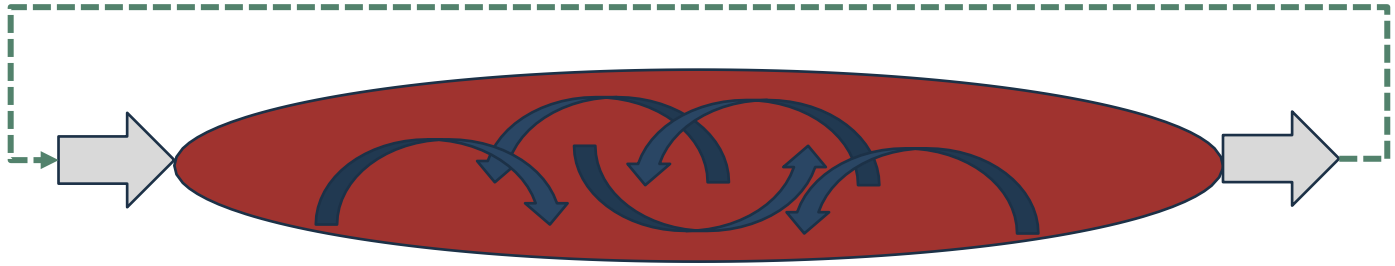
Direct human evaluation
Model tuning by formal proofs

ML: Neural networks/ deep learning



Counterfactual evaluation
Model tuning by back propagation

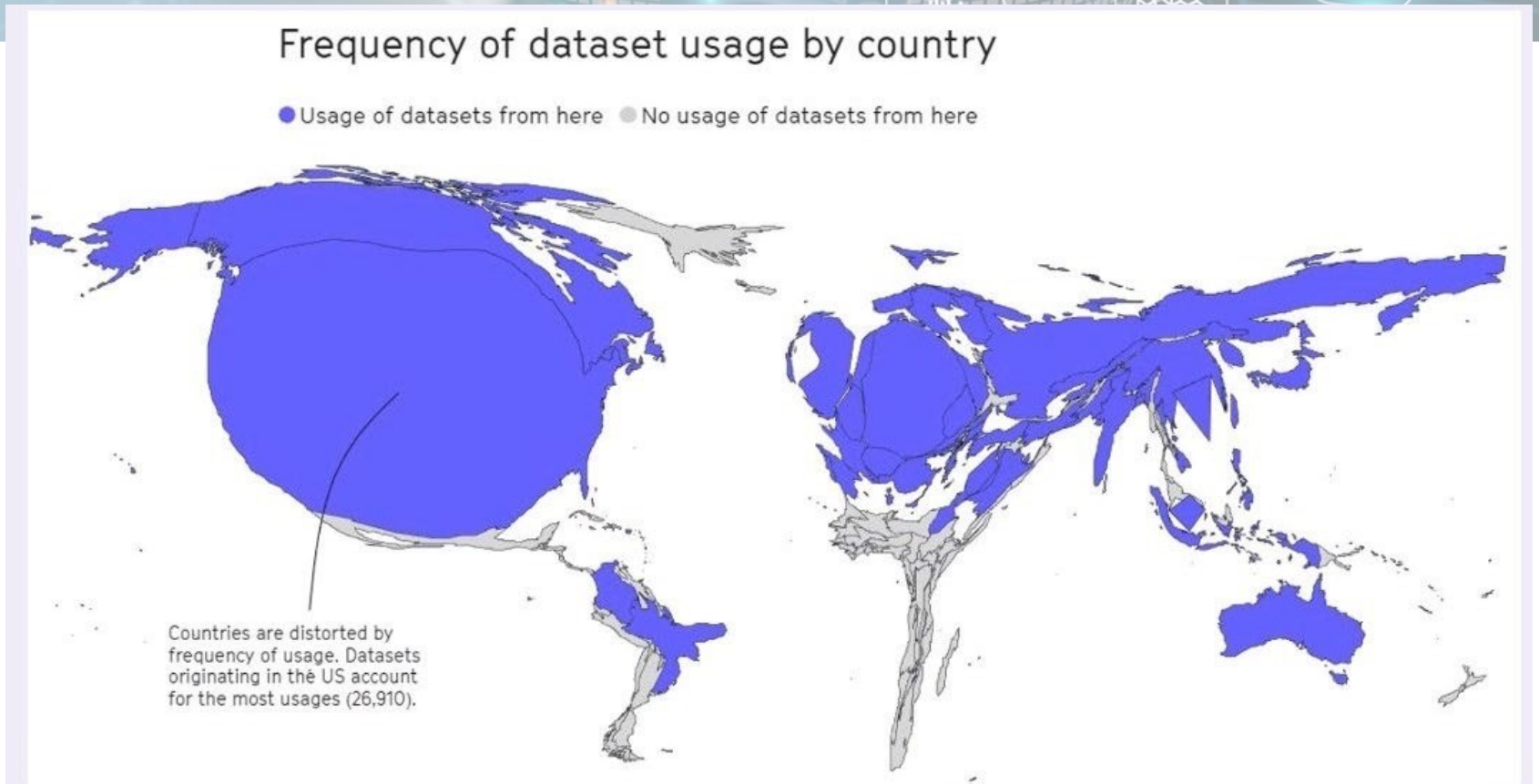
Generative AI/ LLMs



Evaluation: ?
Model tuning: ?

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

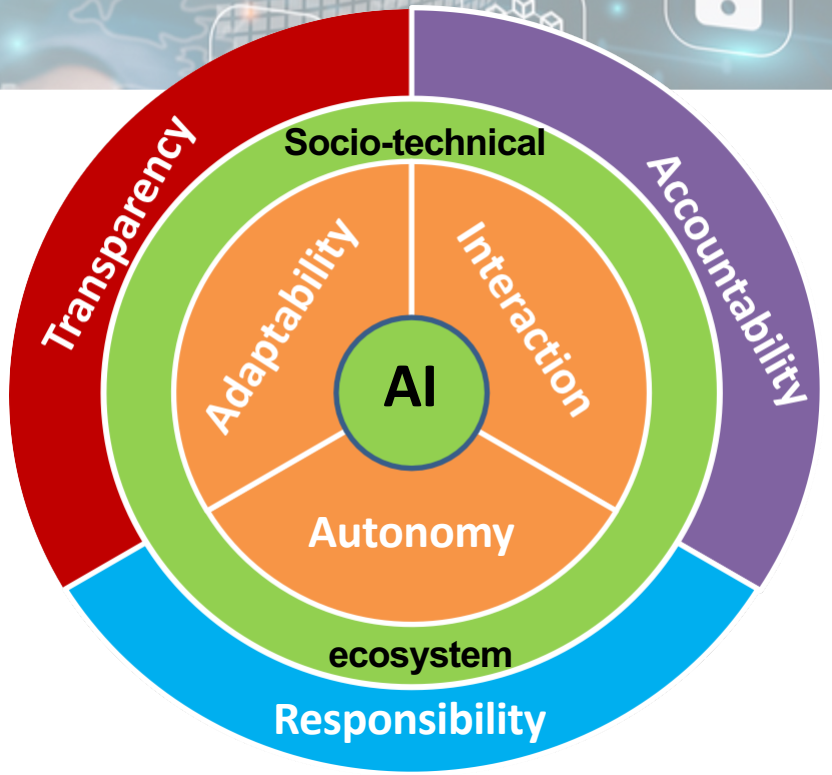
Mozilla internet health report 2022

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
  - …

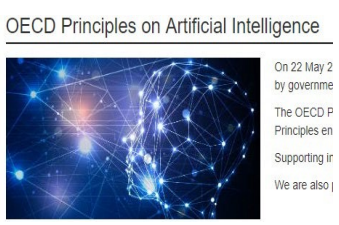
PRINCIPLES AND GUIDELINES

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

EU HLEG	OECD	IEEE EAD
• 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?

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

https://ethicsinaction.ieee.org

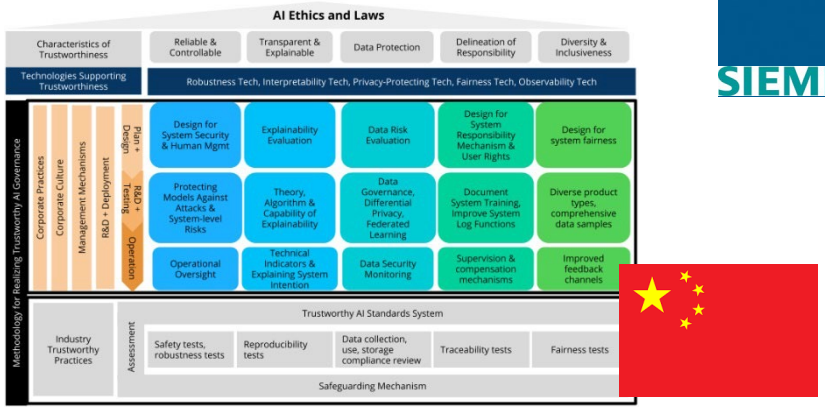


https://www.oecd.org/going-digital/ai/principles/

RESPONSIBLE AI – POLITICS AND BUSI

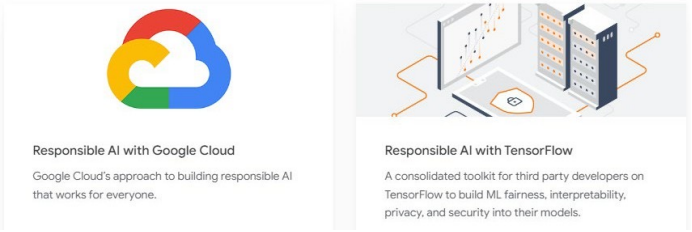
"We need to get in control [of AI] so that we can trust it, and it has human oversight, and – very importantly – that it doesn't have bias"

- Eurocomissaris Vestager



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.



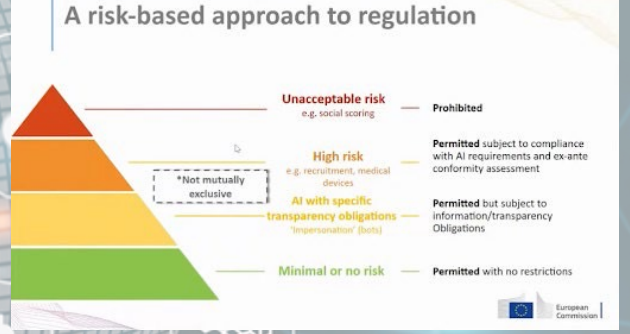
RESPONSIBLE AI IS NOT A CHOICE!

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 differation

AI ACT



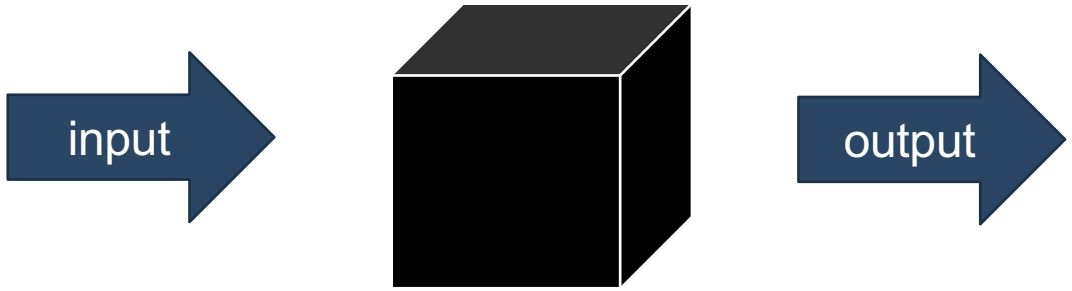
The legislation aims to regulate AI based on its potential to cause harm.

- 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**

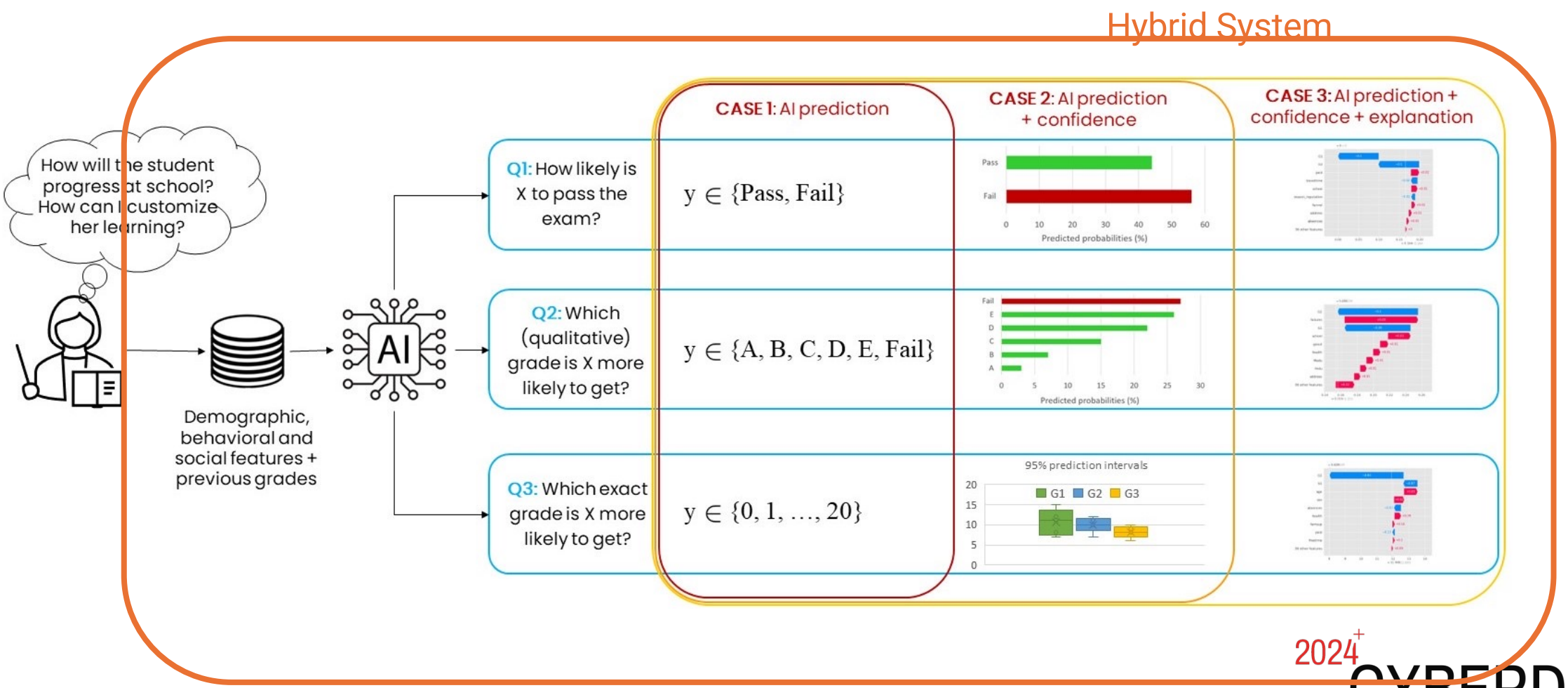
OPERATIONALIZING RAI: ONE PROBLEM



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

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

Explanation empowers human oversight over algorithms



Explanation by counterfactuals empowers what if reasoning

Sorry, your loan application has been rejected.

Our analysis:

The following features were too high:

PercentInstallTrad... | NetFractionRevolv... | NetFractionInstall... | NumRevolvingTra... | NumBank2NatlTra... | PercentTradesWB...

The following features were too low:

MSinceOldestTrad... | AverageMInFile | NumTotalTrades

The following features require changes:

MaxDelq2PublicR... | MaxDelqEver

Counterfactuals suggest where to increase (green, dashed) or decrease (red, striped) each feature.

Counterfactual are based on generative AI

— *Choose one of the case studies we selected using the menu below*

id:156 - class:Melanoma

Image to explain (predicted class)

Melanoma

Neighborhood:

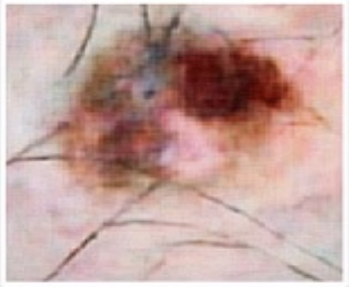
- Melanoma: 7
- Melanocytic nevus: 488
- Basal cell carcinoma: 181
- Actinic keratosis: 32
- Dermatofibroma: 194
- Vascular lesion: 98

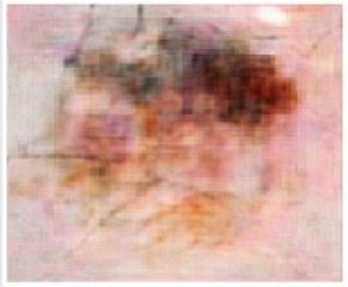
Counter example image (class)

Melanocytic nevus

Prototype images

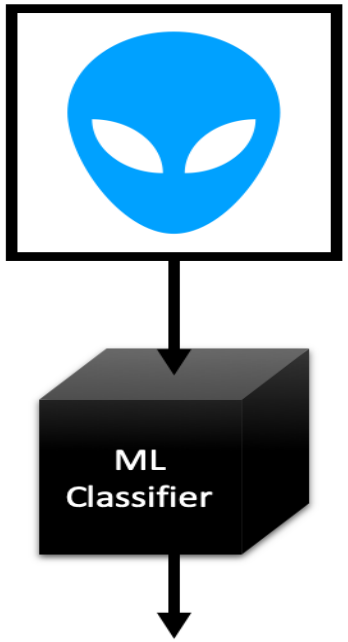
The following images are generated syntethically and they are classified with class **Melanoma** by the blackbox.





Explanation as a Human-Machine Conversation

1



C: I predict: "High Risk of Recidivism"

H: Why?

2

C: See Below

{age < 38 & color = blue & recidivist = True} -> "High Risk of Recidivism"

Because she is younger than 38, has blue color and she is already recidivist

H: *(Hmm. It could be color biased!)* Which training examples are most similar to the instance and influencing the outcome and which not?

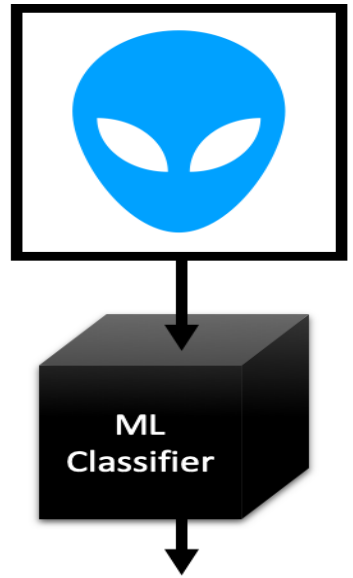
3

C: Similar Instances

C: Dissimilar Instances

H: What happens if you don't consider that she is young and recidivist?

4



C: I still predict "High Risk of Recidivism" because she is blue

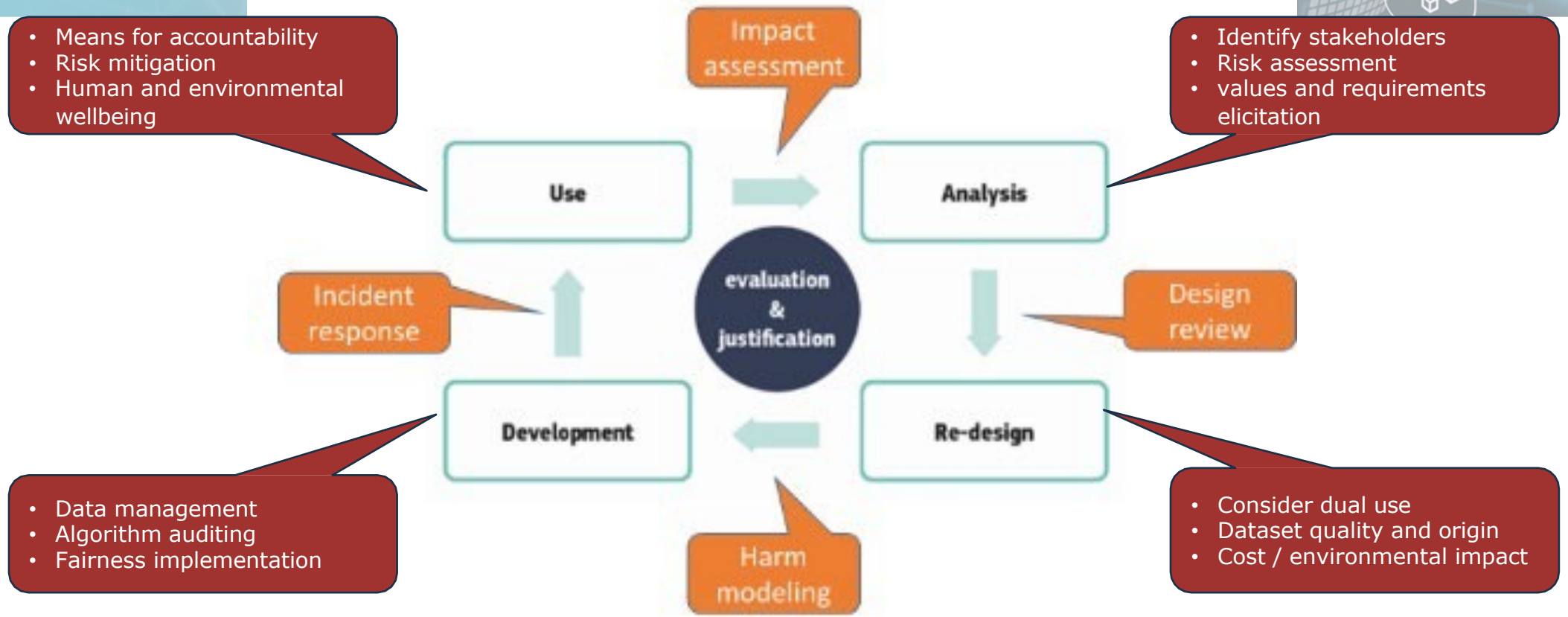
ANOTHER PROBLEM: ALIGNMENT

- 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**!

RESPONSIBLE AI LIFECYCLE



- Means for accountability
- Risk mitigation
- Human and environmental wellbeing

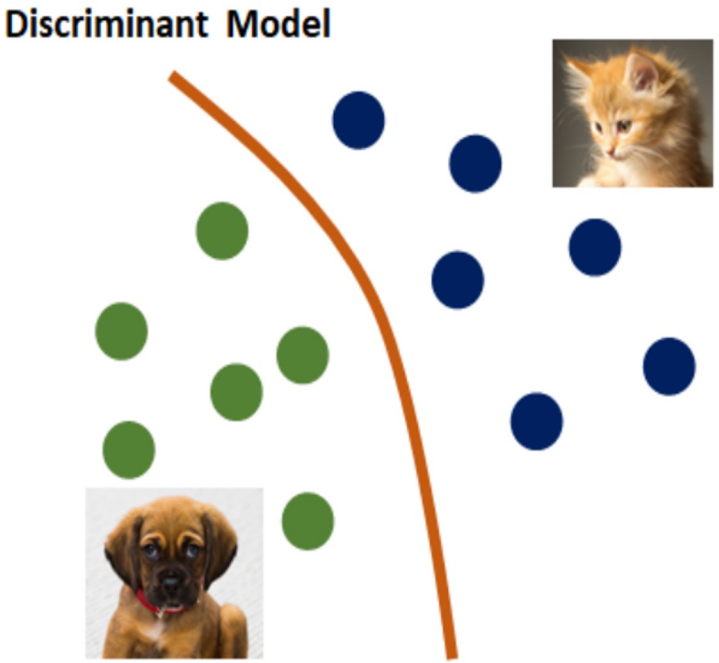
- Identify stakeholders
- Risk assessment
- values and requirements elicitation

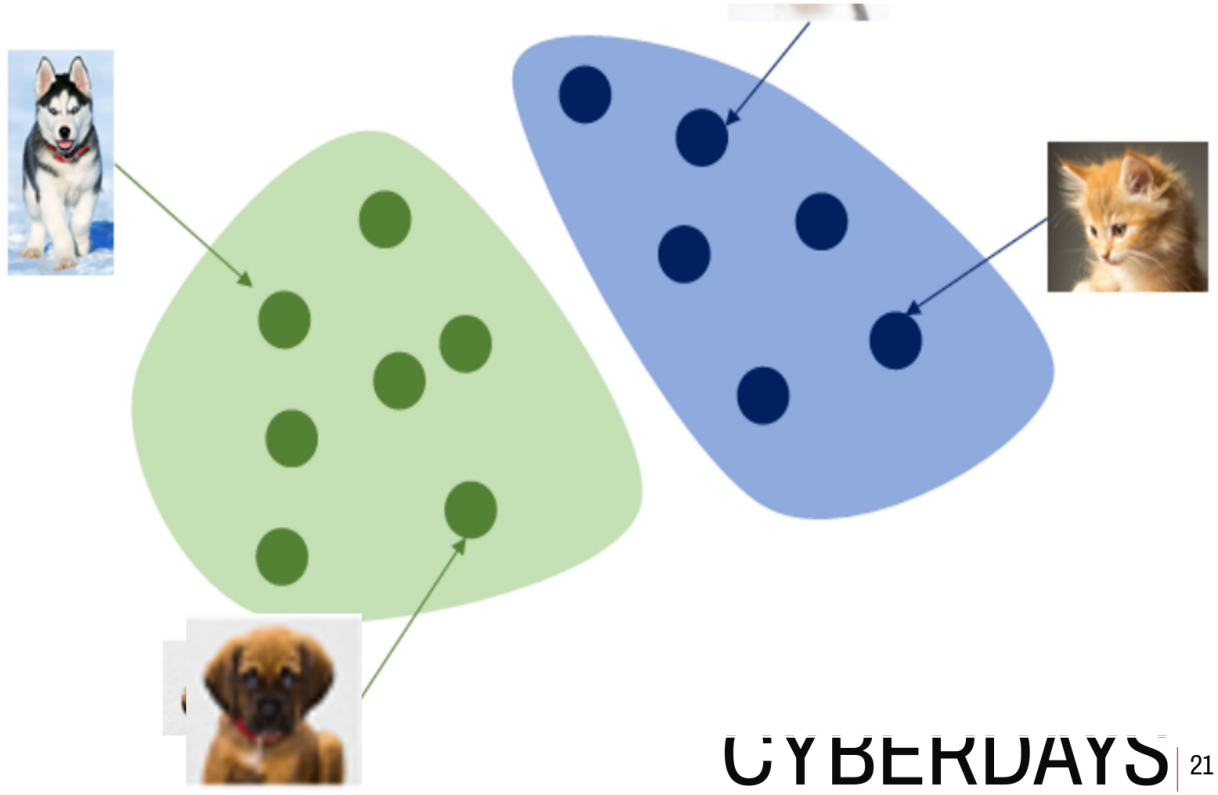
- Data management
- Algorithm auditing
- Fairness implementation

- Consider dual use
- Dataset quality and origin
- Cost / environmental impact

Discriminative vs Generative AI

- **Generative model** capture correlations such as "*things that look like boats are likely to appear near /for images) to things that look like water*" and "*eyes are unlikely to appear on the forehead*".
- **Discriminative models** try to draw boundaries, while generative models try to model the location of data in space.





Example of Generative models

Figure 5: 1024 × 1024 images generated using the CELEBA-HQ dataset. See Appendix F for a

Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." *arXiv preprint arXiv:1710.10196* (2017).

Chu, Mengyu, et al. "Learning temporal coherence via self-supervision for GAN-based video generation." *ACM Transactions on Graphics (TOG)* 39.4 (2020): 75-1.

EXAMPLES OF GENERATIVE AI:

Image Generation

MidJourney
image generation AI

INPUT PROMPTS

- a rabbit wearing a sunhat
- Santa Claus in a blue outfit
- a watercolor of a red bicycle in front of a waterfall

OUTPUT

Real-Life Applications of this Platform
- advertisements, published illustrations, corporate visuals, novel image generation

Text Generation

ChatGPT
text generation AI

INPUT PROMPTS

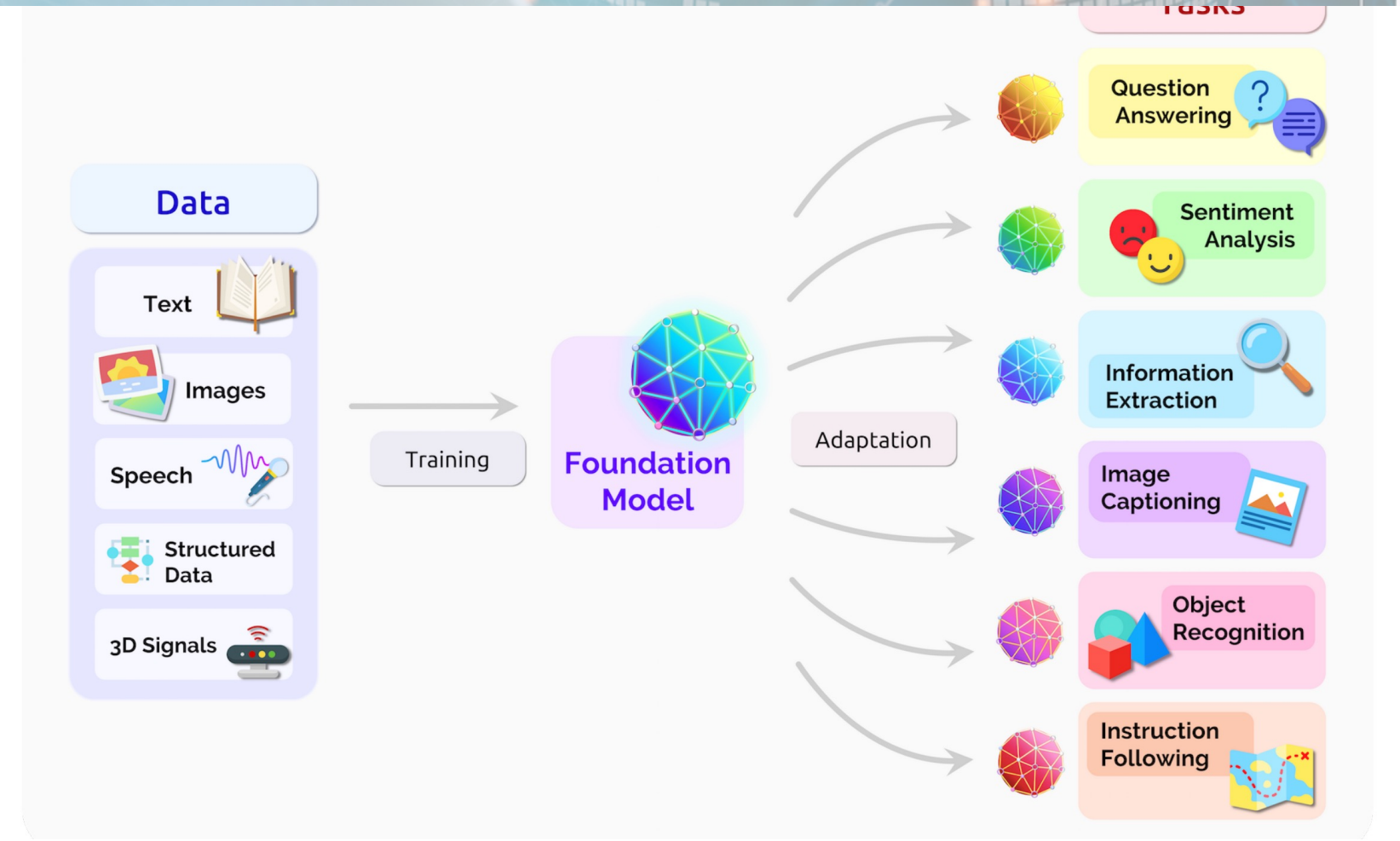
"Write a limerick about data taking over the world. Include the word: egg"

OUTPUT

"Data taking over the world, it's true
It's growing like an egg that's due
It's everywhere we go
It's taking over slow
But it's the future, there's nothing we can do."

Real-Life Applications of this Platform
- communications, journalism, publishing, creative writing, writing assistance

…many tasks--- pieces of decision making



Le trappole

- «Prompting» richiede capacità di fare domande e senso critico. Si parla di i**ngegneria delle domande**. Sta diventando una competenza da acquisire.
- Robustezza: impara **pattern statistici plausibili**, ma non necessariamente corretti: allucinazioni
- Incorporano bias, stereotipi e valori presenti nei dati disponibili che sono per lo più del «nord globale» per lo più inglese e cinese
- Mancanza di multi-lingualità e multi culturalità
- Mancanza di trasparenza sui dati di training per poter capire anche la legalità di tali fonti ed i diritti di proprietà
- Mancano metodi per distinguere tra contenuti generati da umani o dalla macchina: problema della autofagia.
- Aumentare la discriminazione nell'accesso alal conoscenza tra nord e sud del mondo, nonché tra classi sociali di una stessa società

Human or human-like?

- Harari dice: la nostra cultura si basa sul linguaggio, se lo fanno le macchine la nostra civiltà va a rotoli e la fiducia calerà
- Un tema importante sarà distinguere il manufatto della macchina da quello delle persone
- Dove è importante: scuola, lavoro, social media, concorsi etc., ricerca scientifica
- Rimedio: strumentare AI generativa con strumenti che a posteriori permettano i contenuti generati, es. watermarking

2024+ CYBERDAYS | 21 - 22 MARZO

Centro di Competenze 5G

PRISMA

AI la rivoluzione ineluttabil
"non distopia ma utopia possibile"

IN COLLABORAZIONE CON fondazione sistema toscana

INTERNET FESTIVAL FORME DI FUTURO

POR CreO 2014 2020 REGIONE TOSCANA FONDO EUROPEO DI SVILUPPO REGIONALE

FSC Fondo per lo Sviluppo e la Coesione

FUB Fondazione Ugo Bordoni Ricerca e Innovazione

SviluppoToscana S.p.A.

comune di PRATO

UNIONE EUROPEA

REPUBBLICA ITALIANA

Regione Toscana