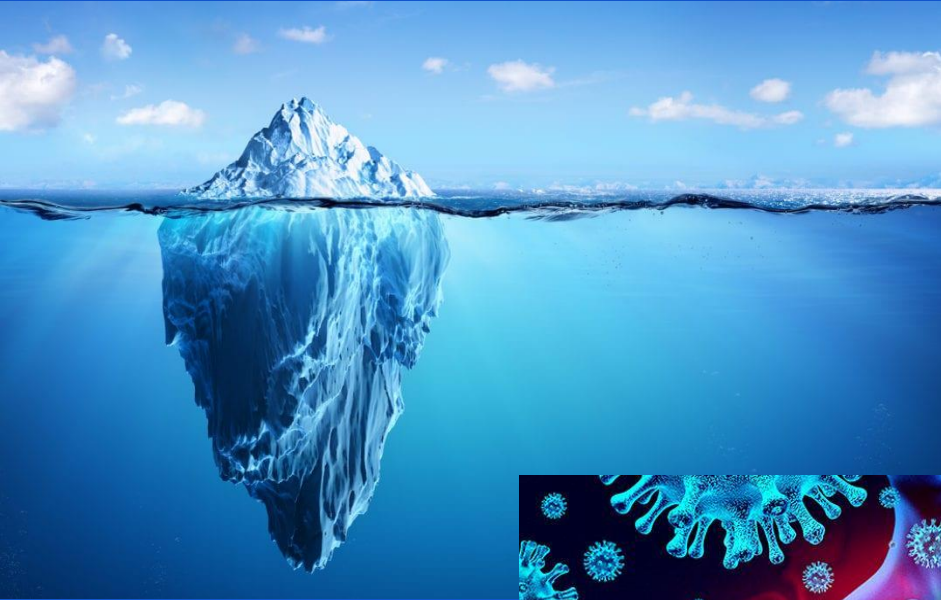


# Artificial Intelligence

*Promise, Perils and Predictions*



# Safety in a Digitalized World



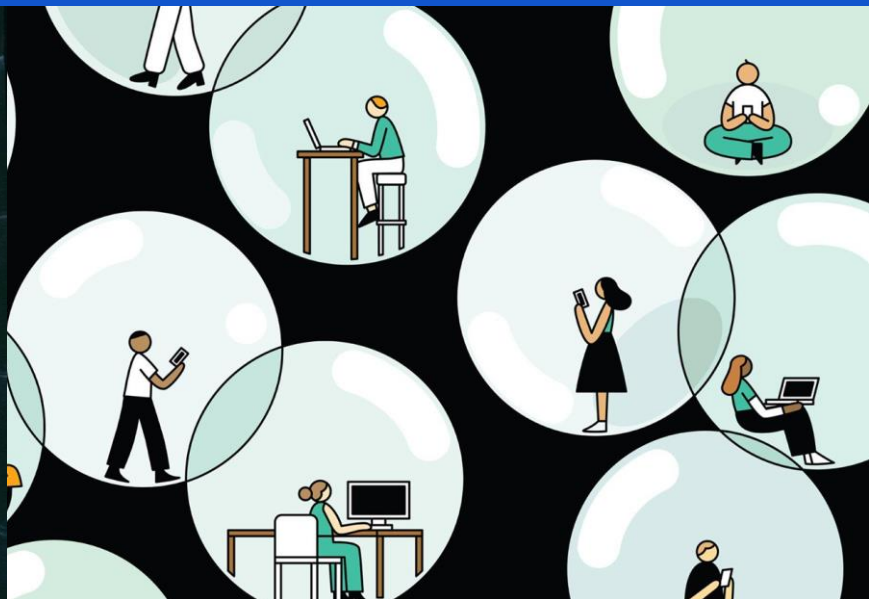






Give the model a prompt, or start telling a story.







Human agency and oversight



Technical Robustness and safety



Privacy and data governance



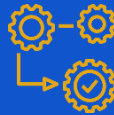
Transparency



Diversity, non-discrimination and fairness



Societal & environmental well-being



Accountability

Lifecycle approach



# Human Agency and Oversight

---

- Data Governance
- Transparency (procedural, strategic and organisational)
- Responsibility
- Consent
- Safety + Security
- Assessments, Audits, Logging Practices
- Human in the Loop
- And many more.







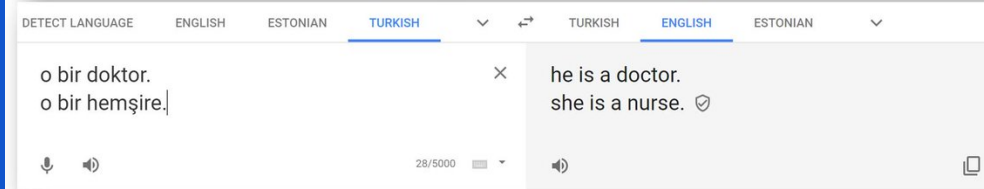
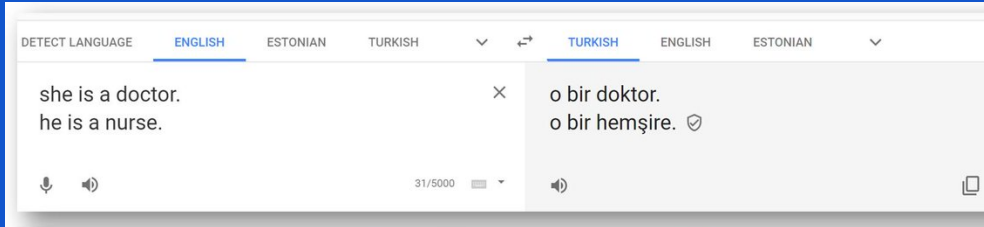
..cascading and catastrophic failure.

Diversity, Non-discrimination and Avoidance of unfair Bias



An AI system's **decisions, choices and suggestions** are not neutral.

# Examples of Bias (historic)



Analyze

Result = happy



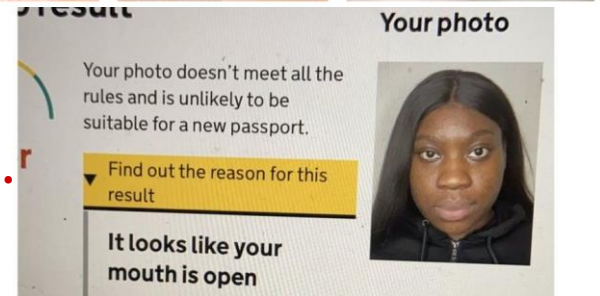
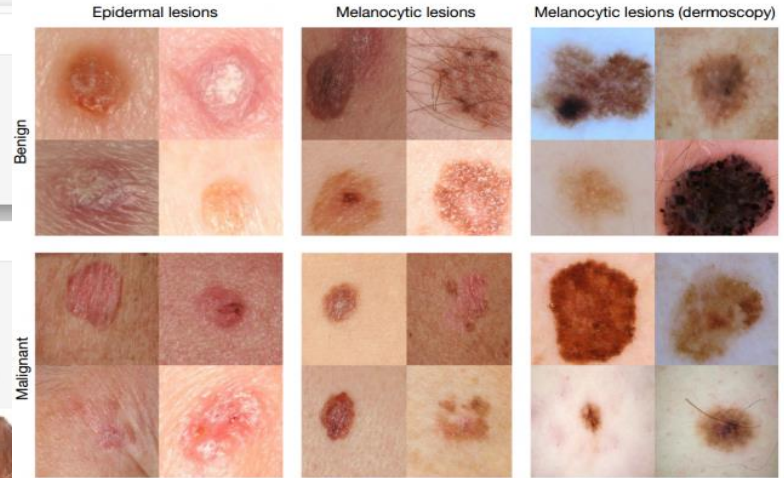
Analyze

Result = afraid



Analyze

Result = happy



Ubiquitous AI systems need to function equally well for all.



AI systems should be technically, safe, accurate, robust and behave as intended.



“panda”

57.7% confidence

+ .007 ×



noise

=



“gibbon”

99.3% confidence

adversarial attack

reward hacking

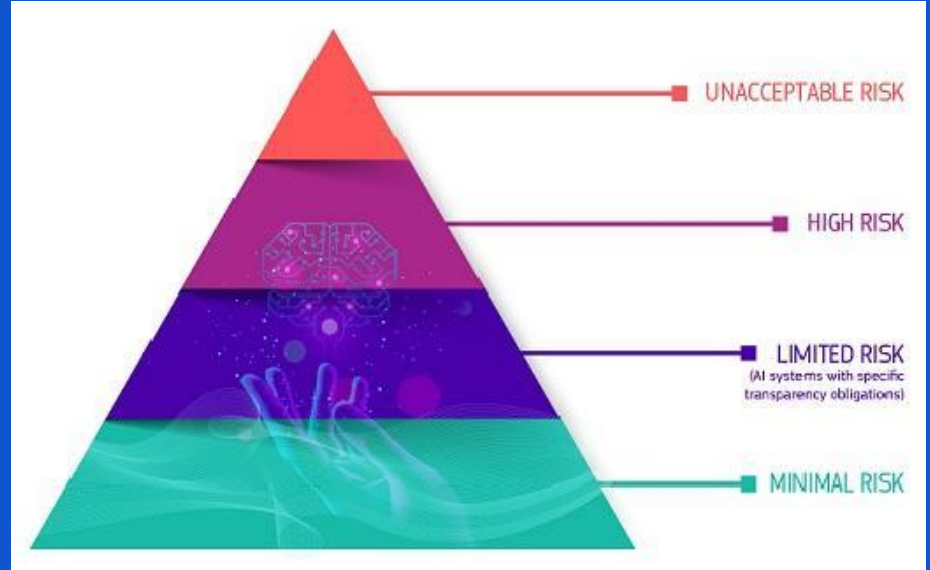




# Geopolitics of AI

---

- ISO standardisation effort for Trustworthy AI
- US-EU Tech and Trade Council for bilateral coordination
- US National Institute of Standards and Technology AI Risk Management Framework
- Australian AI Ethics Framework  
European Commission's AI Act



# From ethical considerations to legal implications: High risk AI systems

---

- AI systems in Annex III: “high risk” – ‘stand-alone high risk AI systems’
  - Biometric identification and categorisation of natural persons
  - Management and operation of critical infrastructure
  - Education and vocational training
  - Employment, workers management and access to self-employment
  - Access to and enjoyment of essential private services and public services and benefits
  - Law enforcement
  - Migration, asylum and border control management
  - Administration of justice and democratic process

Conformity assessment



- Control
- Acceptance
- Understanding
- Prioritise technology that adapts to us and not vice-versa

**Let's shape our future and not be shaped by it.**



# Thank you.

Andonix

© Effective State and Inclusive Development

© Harvard medical School

© DeviantArt VSFX

© Autonomous vehicle international

© Venafi

© Vox.com with GPT-2, OpenAI

© DALLÉ-2, OpenAI, the Verge

© 1997 Rex Features

© NBC news

© Icons made by Freepik from [www.flaticon.com](http://www.flaticon.com)

© NY Post

© Bettal.uk

© Ebuyer.com

© 3BL Media

© Ebuyer.com

© 3BL Media

© Rachel Thoma

© B. Nalini

© Esteva, Kuprel et. al, 2017

© Elaine Owusu

© Explaining and Harnessing Adversarial Examples,  
Goodfellow et al, ICLR 2015

© OpenAI, J Clark & D Amodei, 2016

© European Commission, Digital Strategy

© Apple

© Warable

© Tavvy