Ethical Challenges of AI

Laurence Brooks*, Professor of Information Systems

Information School, University of Sheffield

IADIS International Conference Information Systems 2025

Madeira Island, Portugal

*l.brooks@sheffield.ac.uk

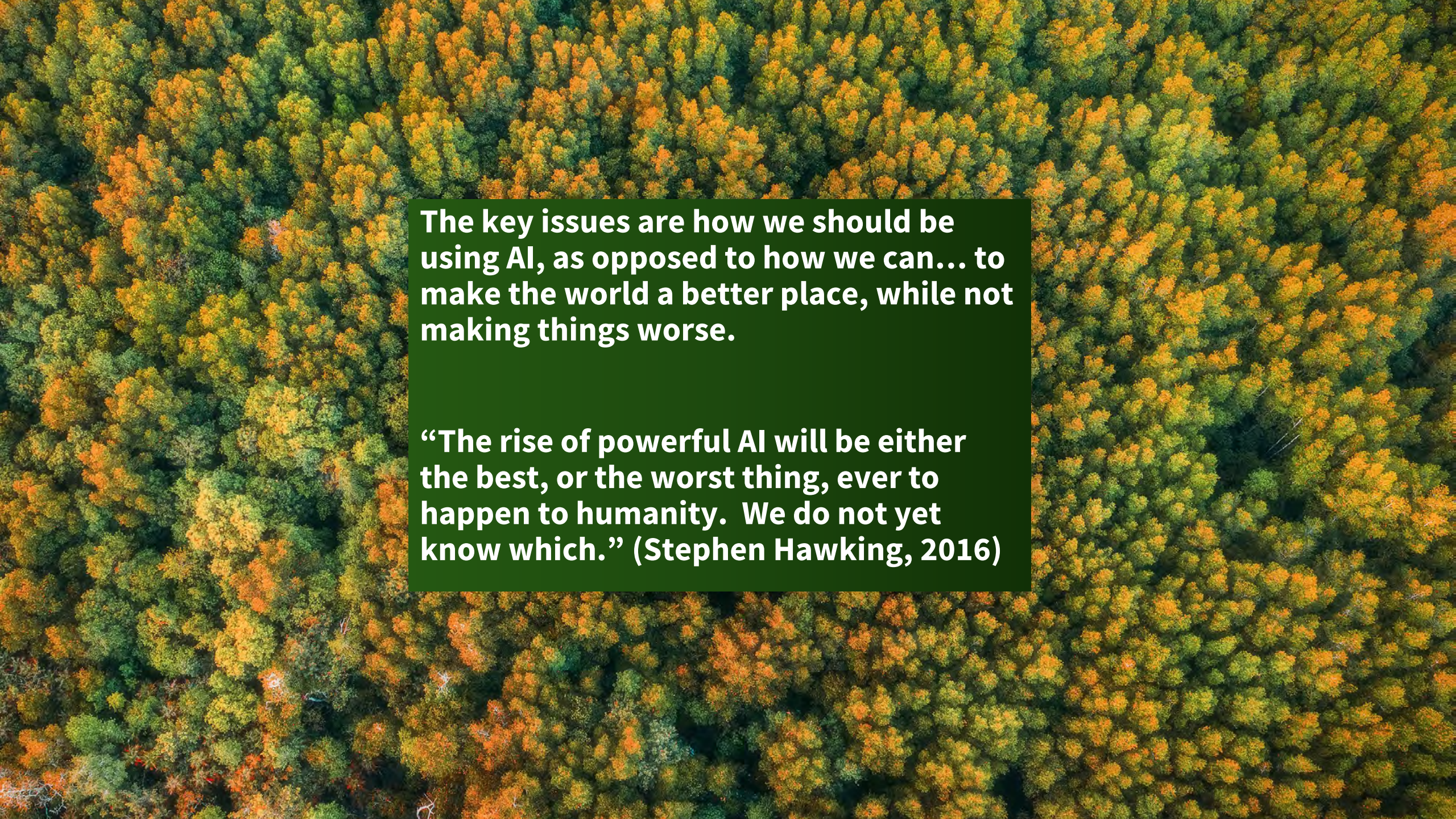
University of Sheffield

A World TOP 100 University

WHAT UNI? STUDENT CHOICE Awards 2024 UNIVERSITY OF THE YEAR

Overview

- As we reach nearly a quarter of the first century in this new millennia, we can see that technology, in all its facets, plays an integral part in all our lives.
- From the smart technology in our pockets to the networks spanning the globe that enable those devices to work anywhere at any time, to the potential of smart technologies, such as AI, to enhance our lives through healthcare, education, finance and so much more.
- But what prices are we paying?
  - Do we really know what the various implications of those technologies are, from the power which is embedded within them by the choices developers make to the opportunities afforded, or not, by the choices the deployers make.
- This talk will explore the various ethical and responsibility issues around technology, not just for today but for the near future, and possibly, further future…



The key issues are how we should be using AI, as opposed to how we can… to make the world a better place, while not making things worse.

“The rise of powerful AI will be either the best, or the worst thing, ever to happen to humanity. We do not yet know which.” (Stephen Hawking, 2016)

Why is this important?

- News story “Passport facial recognition checks fail to work with dark skin” (October 2019)
  - Home Office passport checking service maps a person’s facial features from computer scanned image/live feed;
  - Compares with previously created facial maps database to find match;
  - Trouble matching the faces of some ethnic minorities.
  - Home Office documents released under a Freedom of Information (FOI) request show it knew its passport photo system failed to work well for some ethnic minority people but decided to use it anyway.
  - Bias in the database, or something else?
- While not new, the issue of AI bias still makes the news,
  - April 2022, the Associated Press reported on an algorithm being used to help social workers in the US decide which families to investigate because of possible child neglect, which seems to harden racial disparities, and if used alone would have flagged a disproportionate number of black children compared with white children.
- January 2025 - DeepSeek repeated false claims **30%** of the time and provided non-answers **53%** of the time, resulting in an **83%** fail rate (https://www.newsguardrealitycheck.com/p/deepseek-debuts-with-83-percent-fail).

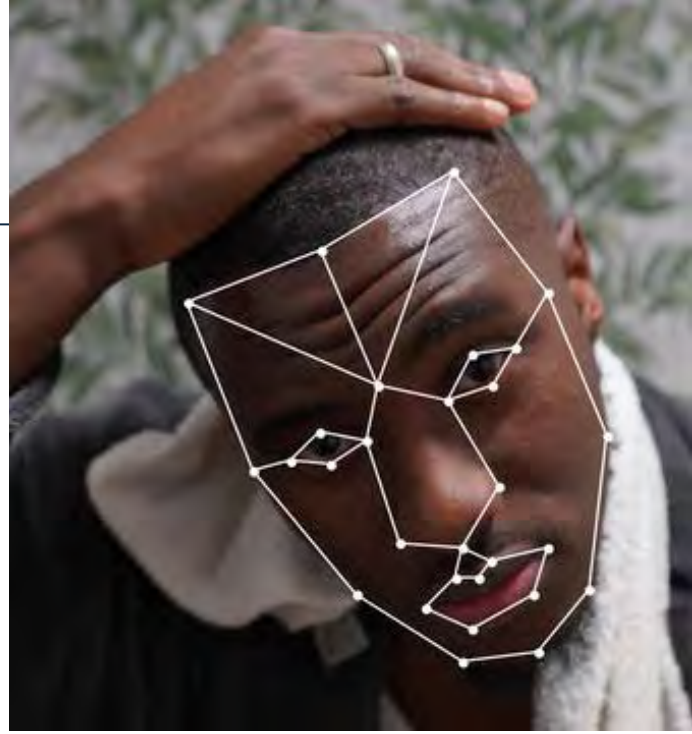
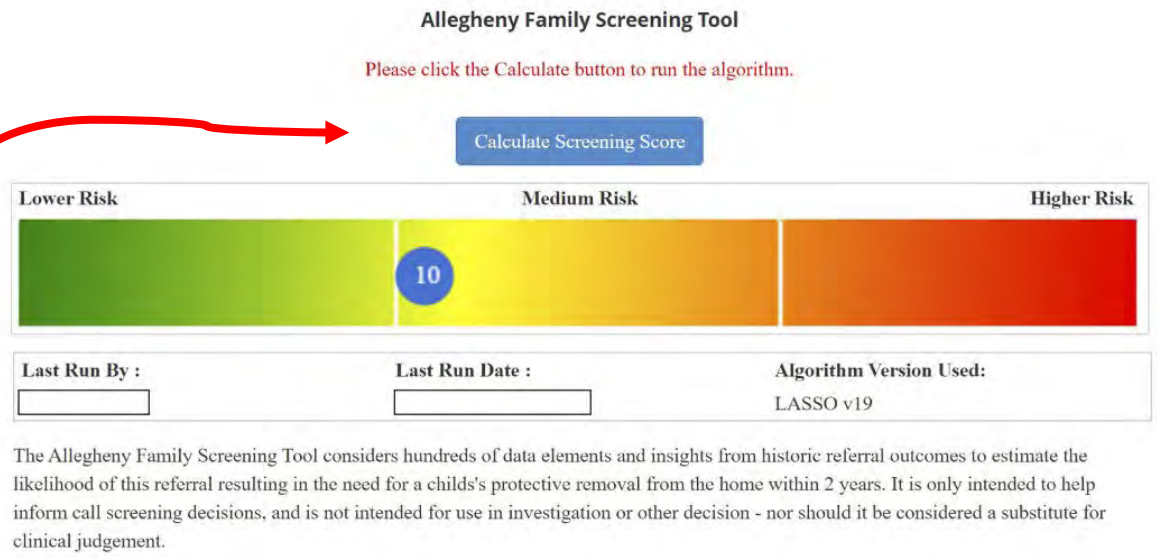


Image by Comuzi / © BBC / Better Images of AI / Mirror B / CC-BY 4.0



The issues

Artificial Intelligence (AI) continues to see heavy investments from industry and governments around the world. While some envision AI as a way to empower individuals and improve society, it is increasingly clear that the ethical ramifications of AI systems and their impact on human societies requires deep and urgent reflection. How can we chart a course through this new AI enabled landscape, which supports everyone?

Jamillah Knowles & Reset.Tech Australia / Better Images of AI / People on phones (portrait) / CC-BY 4.0

- At a recent conference, these are some of the questions that were envisioned in this area:
  - How should AI systems talk to humans?
  - What effects do gendered persona have on human perceptions of AI?
  - Should AI be allowed to use methods of nudging and persuasion to influence people?
  - What is the human perception of AI agency and what does that imply for moral agency?
  - How can AI support human autonomy?
  - To what extent does AI need to be able to explain its decisions in a way humans understand?
  - How do we know that an AI system is trustworthy?
  - How should AI systems interact with groups of humans (e.g., in the context of teams such as the police force, the military etc.)?
  - What are the ethical concerns related to AI as bosses in a work context?
  - What are the cultural differences related to human-AI interactions and how to address these in design and/or governance?

AI definitions (some, there are lots)

“AI is a collection of technologies that combine data, algorithms and computing power” (European Commission, White Paper, 2020)

“An AI system is a machine-based system that, for a given set of explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment”. (OECD, 2023)

AI system as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" (Art. 3(1) EU AI Act, 12 July 2024)


EUROPEAN COMMISSION

Brussels, 19.2.2020
COM(2020) 65 final

WHITE PAPER

On Artificial Intelligence - A European approach to excellence and trust

EUROPEAN COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposal for a

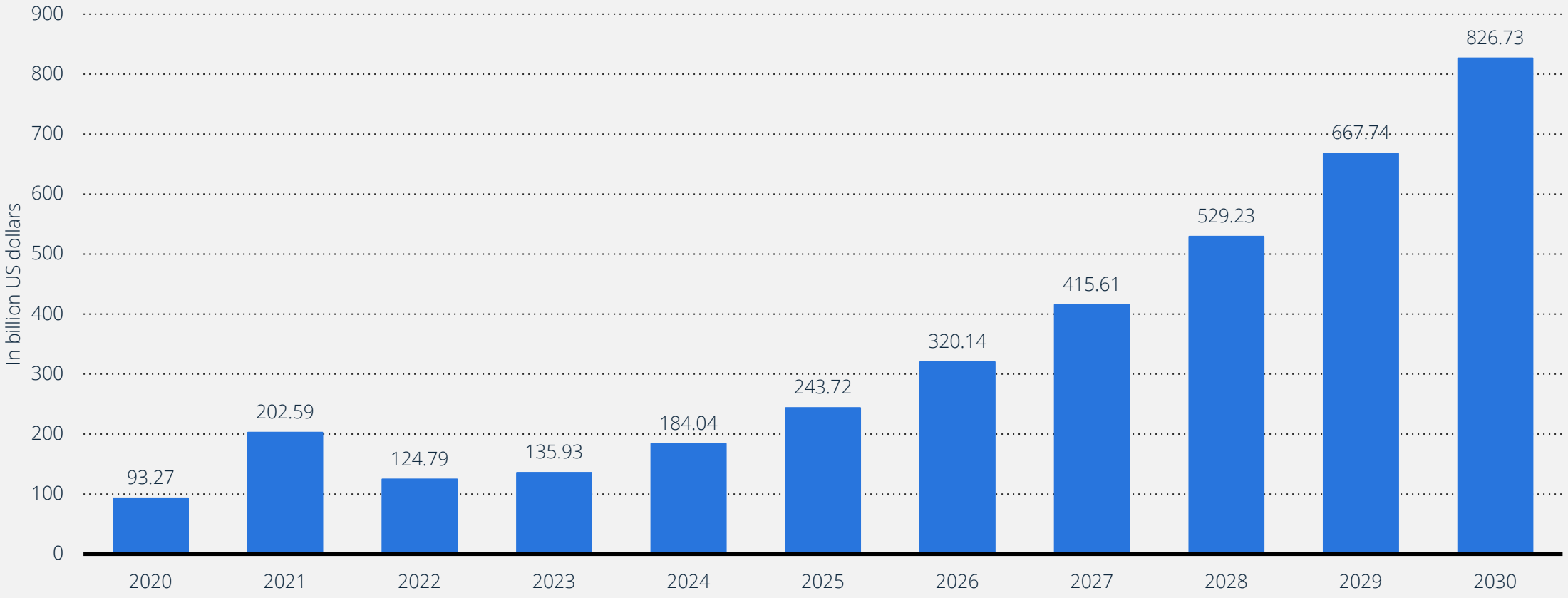
REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

Artificial intelligence (AI) market size worldwide from 2020 to 2030 (billion US dollars)

AI market size worldwide from 2020-2030



Source(s): Statista; ID 1474143

Types of AI based on capabilities

Artificial General Intelligence (AGI)
A type of AI that could learn to accomplish any intellectual task that humans can perform or surpass human capabilities in many economically valuable tasks. While AGI is often mentioned in the news and on social media, there is currently no research that proves how AGI could be developed or achieved.

Artificial Narrow Intelligence (ANI)
Also known as weak AI or narrow AI, is designed to perform a specific set of tasks. All AI in existence today is narrow AI, usually using machine learning or deep learning techniques. Examples of narrow AI include internet search engines, recommendation systems and facial recognition. Such AI tools are designed to perform tasks within a single, defined set of problems.

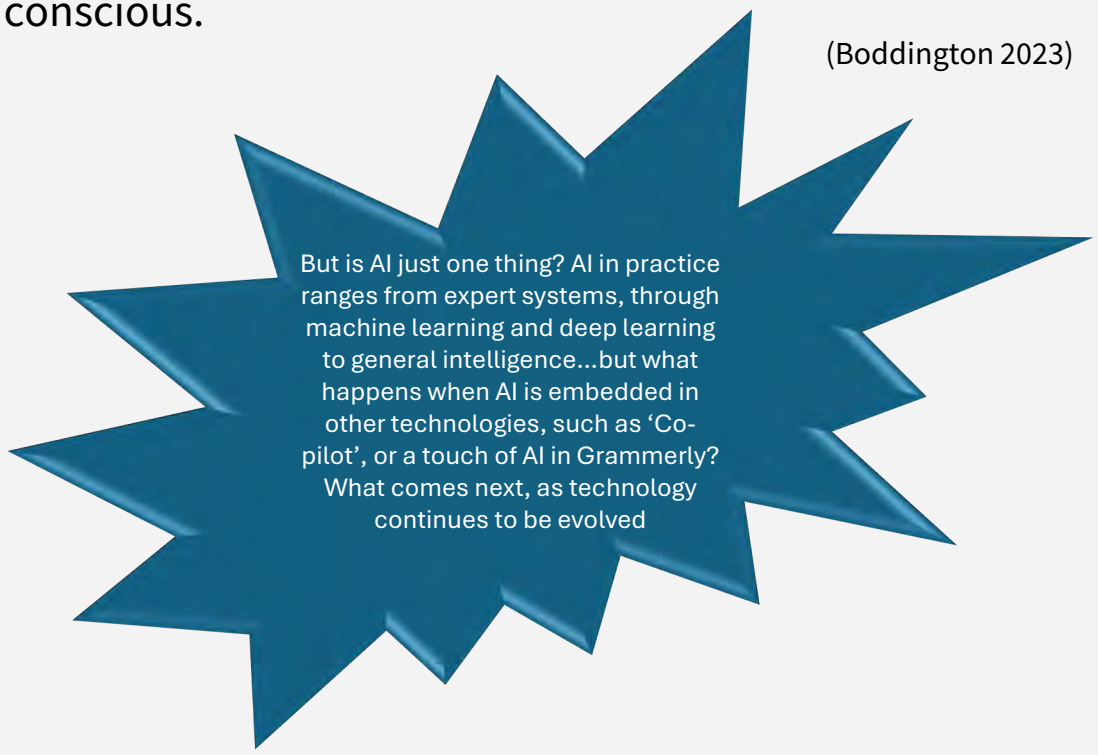
Artificial Superintelligence (ASI)
Is a speculative concept of AI that would far surpass human intelligence, exceeding in memory, data-processing and decision-making abilities.

Scottish AI Alliance: https://www.scottishai.com/news/the-different-types-of-artificial-intelligence

Weak AI: the attempt to build programmes that demonstrate capabilities of intelligence, without necessarily being 'intelligent' themselves.

Strong AI: the attempt to build programmes that have intelligence in the form of understanding and/or that are conscious.

(Boddington 2023)

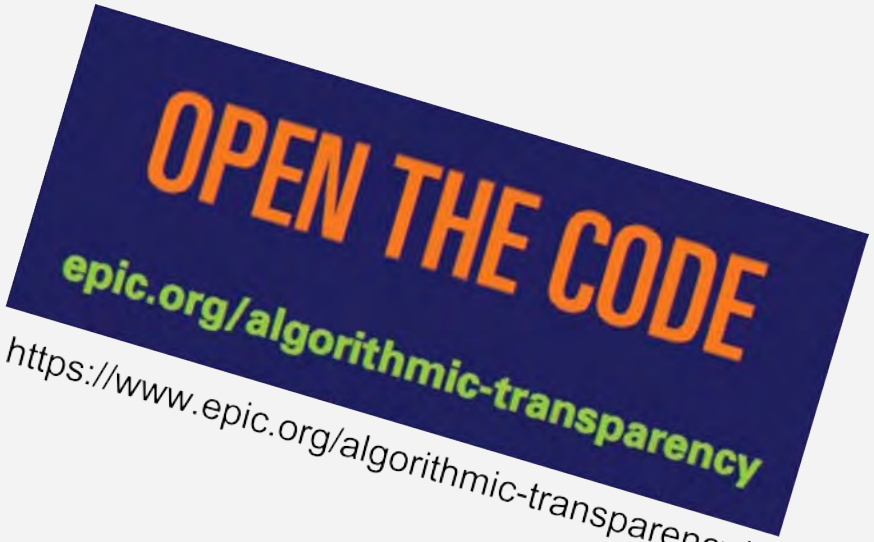


Some concerns…

https://www.canterbury-cathedral.org/about/privacy/

https://www.technewsworld.com/story/85544.html

https://www.internationalworkplace.com/news/landmark-case-broadens-discrimination-protection-for-cancer-victims-55964



https://www.epic.org/algorithmic-transparency/

Artificial intelligence + Add to myFT

Political deepfakes top list of malicious AI use, DeepMind finds

Artificial intelligence is used more to create realistic but fake celebrity images than to assist cyber attacks, Google unit says

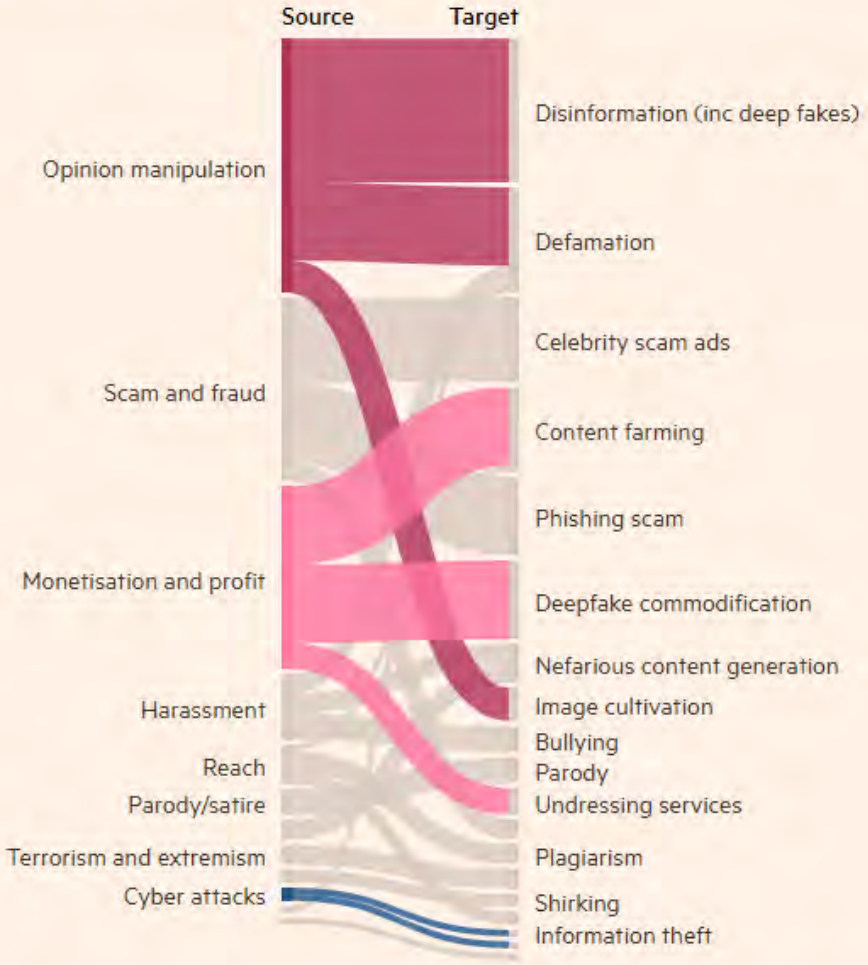
Deepfakes of UK Prime Minister Rishi Sunak have appeared on TikTok and Instagram ahead of next week's general election © Leon Neal/AFP/Getty Images

Cristina Criddle in London YESTERDAY

18

Generative AI is most commonly used to create deepfakes and other disinformation

Motivations of bad actors linked to techniques



FINANCIAL TIMES

Source: Jigsaw, DeepMind

FINANCIAL TIMES

Generative

WIRED SECURITY POLITICS GEAR BACKCHANNEL BUSINESS SCIENCE CULTURE IDEAS MERCH

BY DAVID GILBERT POLITICS JUN 28, 2024 5:00 AM

Neo-Nazis Are All-In on AI

Extremists are developing their own hateful AIs to supercharge radicalization and fundraising—and are now using the tech to make weapon blueprint bombs. And it's going to get worse.

Deepfakes of UK Prime Minister Rishi Sunak have appeared on TikTok and Instagram ahead of next week's general election © Leon Neal/AFP/Getty Images

Cristina Criddle in London YESTERDAY

18

FINANCIAL TIMES

VATICAN NEWS POPE VATICAN CHURCH WORLD

Radio Vaticana ROSARY IN LATIN

Pope Francis addresses Minerva Dialogues (Vatican Media)

POPE POPE FRANCIS SCIENCE AND TECHNOLOGY AUDIENCES

Pope Francis urges ethical use of artificial intelligence

While praising the benefits of technology and artificial intelligence, Pope Francis says AI raises serious questions and must be ethically and responsibly used to promote human dignity and the common good.

By Deborah Castellano Lubov

Source: Jigsaw, DeepMind

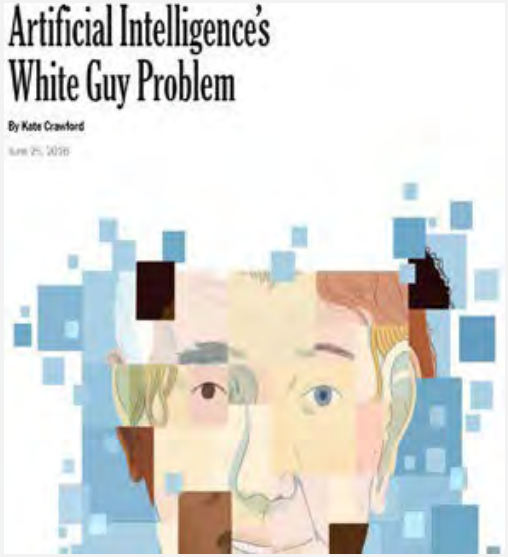
Meet Larissa Wagner: She is an attractive young woman from Germany whose early Instagram posts display the usual pics of her on hikes or chilling at home.

But: Turns out Wagner is a big fan of Germany's far-right AfD party, something she started relentlessly posting about on her X and Insta accounts as polling day approached. **Oh, and Wagner isn't real**.

Dystopian: She's one of many AI influencers created to farm clicks, sell products ... or influence elections. And while her Instagram bio states she's an AI model, it's clear from the flood of comments she receives that hardly anyone has noticed or properly understands what that means.

Nightmare fuel: Wagner's exposure in the media presumably dampened her reach, but responding to Sky News, it's clear "her" creator isn't phased. "I think it's completely irrelevant who controls me," they (it?) said. "Influencers like me are the future."

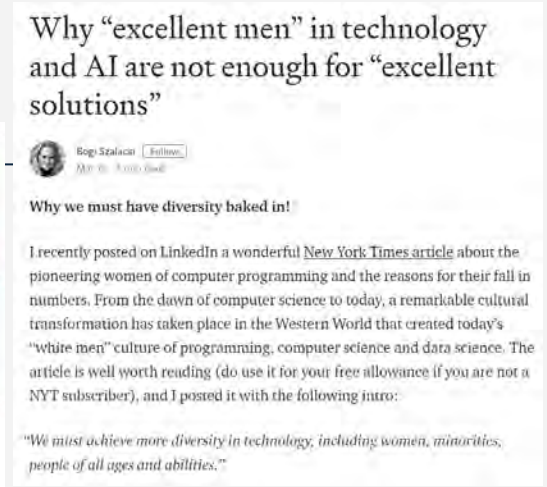
Diversity & gender

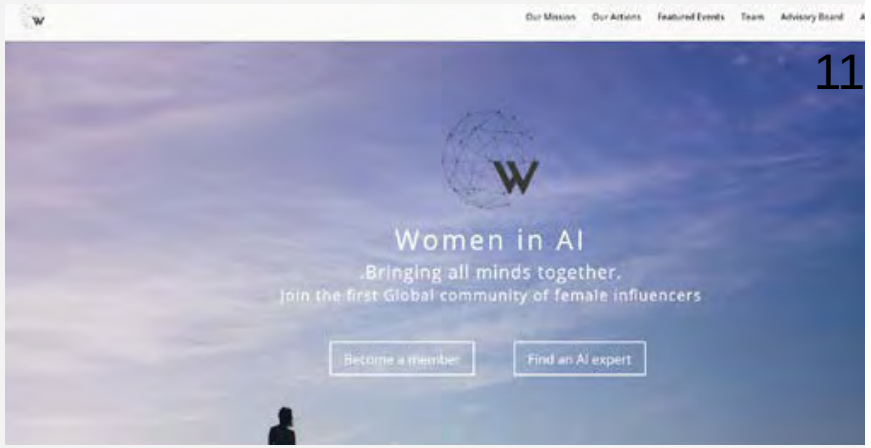


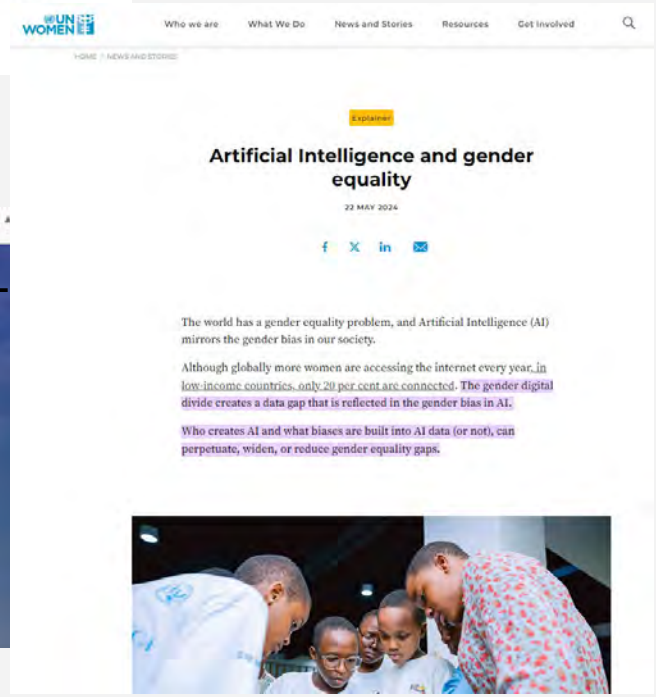
Artificial Intelligence and Robotics

AI has a gender problem. Here's what to do about it

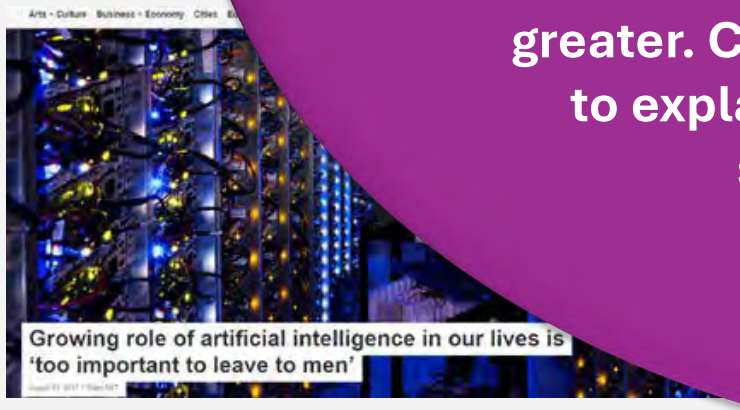
Submissive female robots, servile voice assistants - does AI need a feminist revolution?

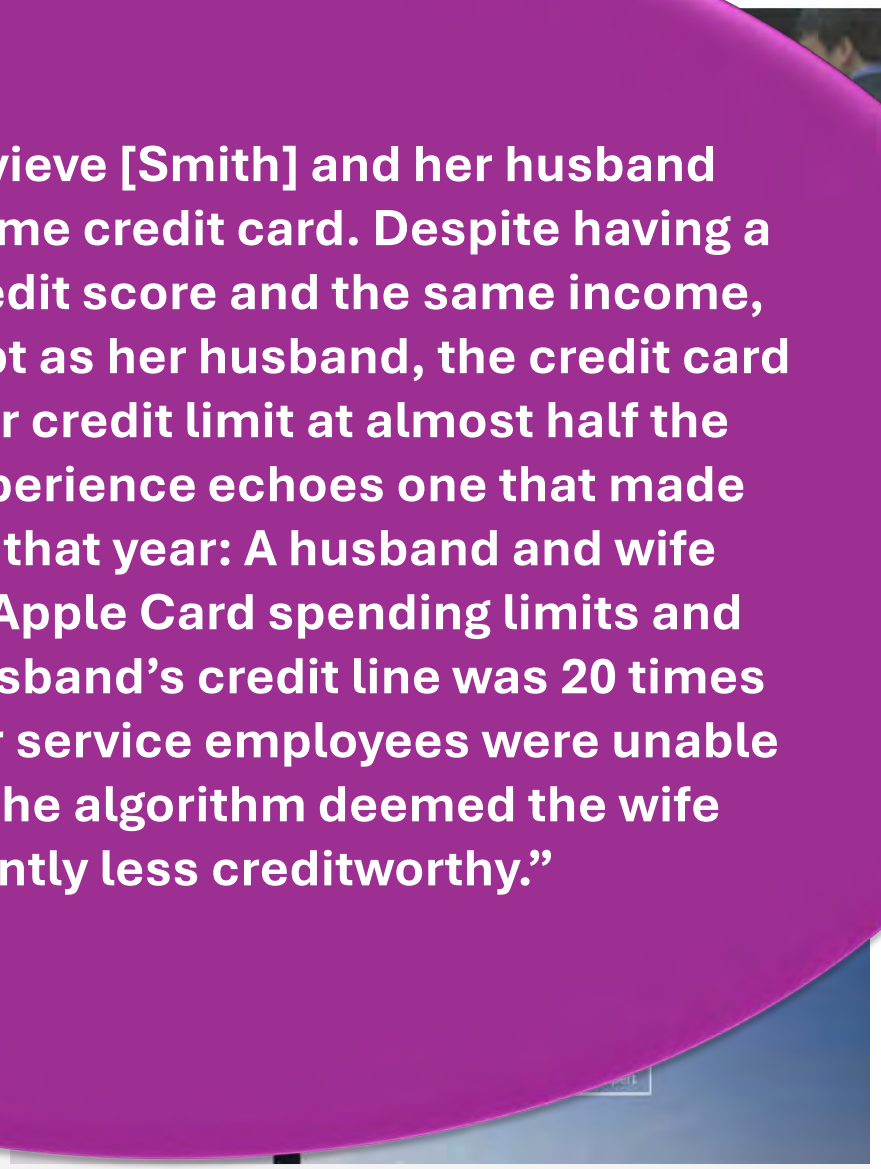


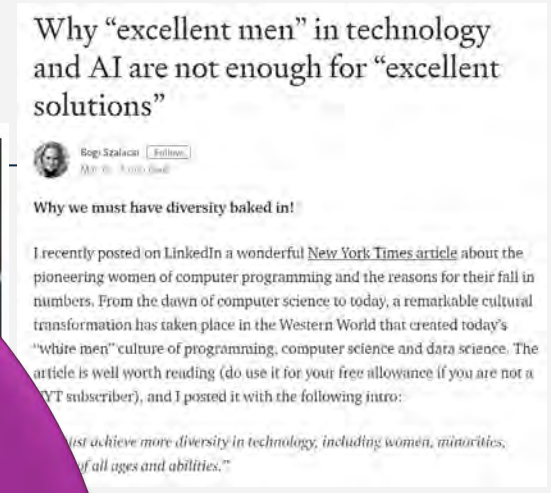


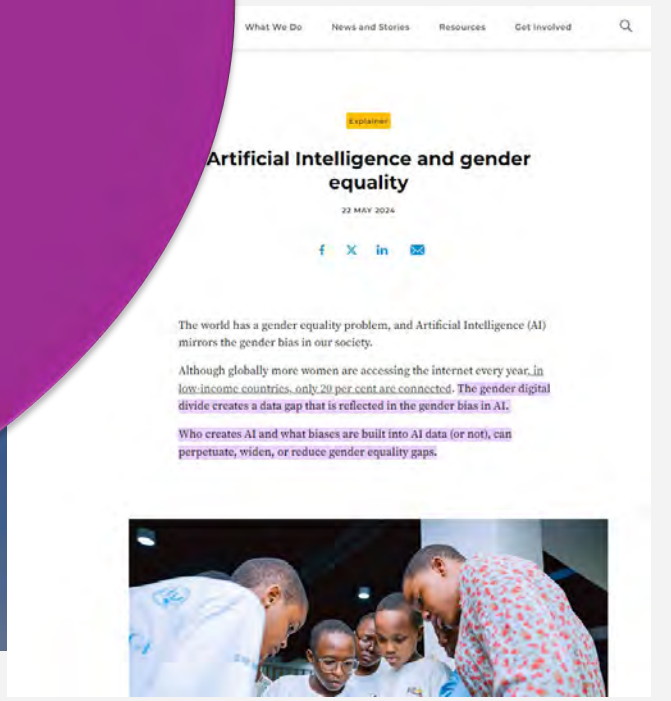


Diversity & gen







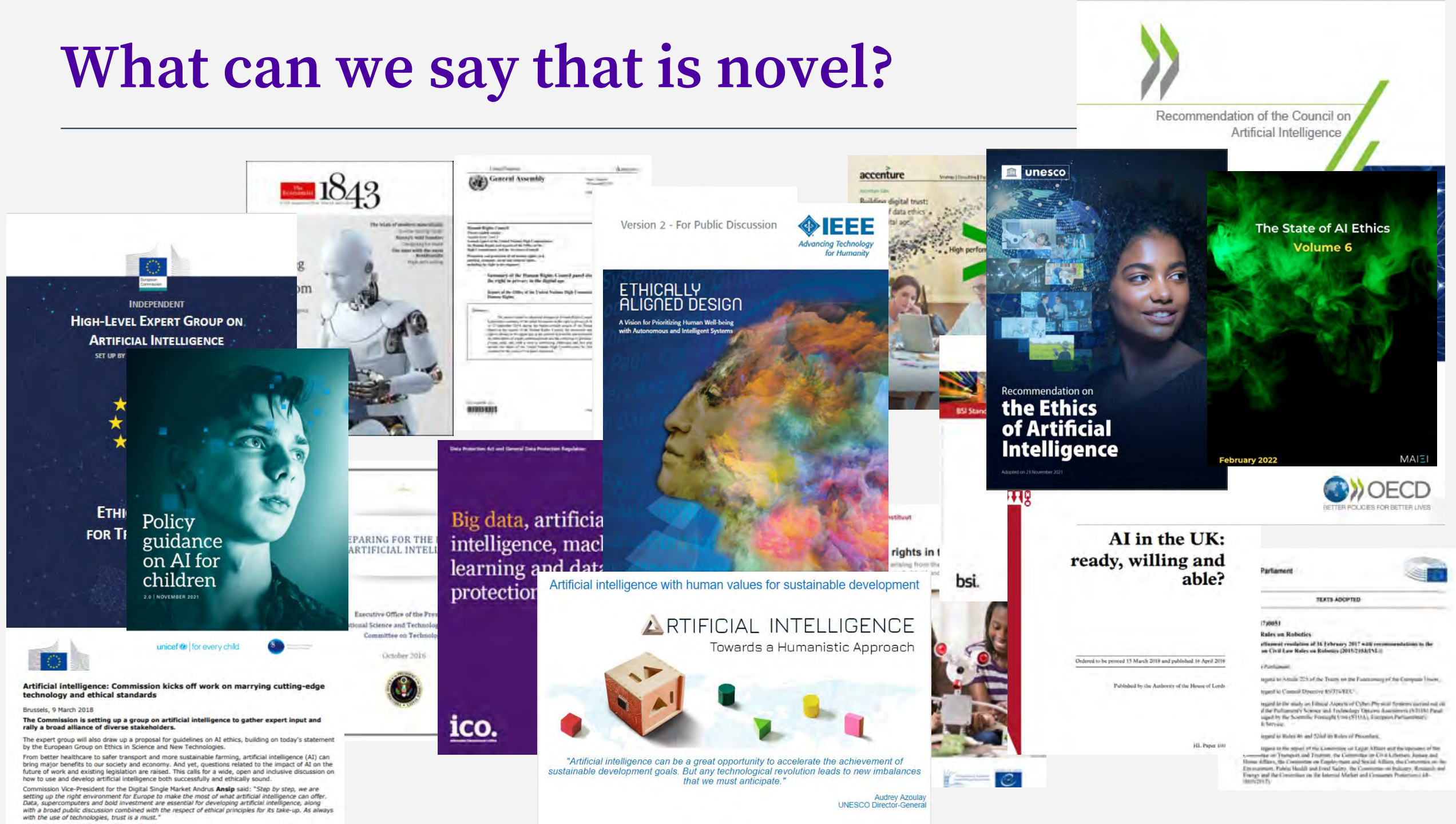


"In 2019, Genevieve [Smith] and her husband applied for the same credit card. Despite having a slightly better credit score and the same income, expenses, and debt as her husband, the credit card company set her credit limit at almost half the amount. This experience echoes one that made headlines later that year: A husband and wife compared their Apple Card spending limits and found that the husband's credit line was 20 times greater. Customer service employees were unable to explain why the algorithm deemed the wife significantly less creditworthy."

Need for responsible AI

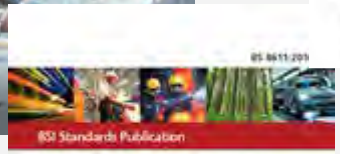
- **Air Canada's Chatbot Misinformation:** A significant legal setback occurred when Air Canada's chatbot provided incorrect airfare information to a traveller, leading to a lawsuit after the airline refused a refund.
    - This case exemplifies the tangible risks to businesses when AI systems malfunction, impacting both financials and reputation.
- **Google's Gemini Controversy:** Google faced public backlash when its Gemini model inaccurately depicted historical images, prompting the company to suspend the AI's image generation feature.
    - This incident underscores the importance of accuracy and accountability in AI-driven content generation.
- **Apple's Siri Voice Controversy:** A UNESCO study highlighted gender bias in Siri and similar voice assistants, prompting Apple to introduce more vocal options for users.
    - This move towards inclusivity demonstrates the broader societal impact of AI and the need for diversity in AI voice interfaces.
- **ChatGPT Regulatory Challenges:** OpenAI encountered regulatory hurdles with ChatGPT in Italy due to the lack of a legal basis for data collection and processing, alongside the absence of an age-verification mechanism.
    - This incident highlights the crucial role of compliance and privacy considerations in AI deployment.
- **Microsoft Bing Chat's Behavioural Anomalies:** Users reported instances of Bing Chat providing incorrect information and exhibiting unexpected emotional responses.
    - These issues spotlight the complexities of AI behaviour and the need for ongoing monitoring and refinement.

What can we say that is novel?

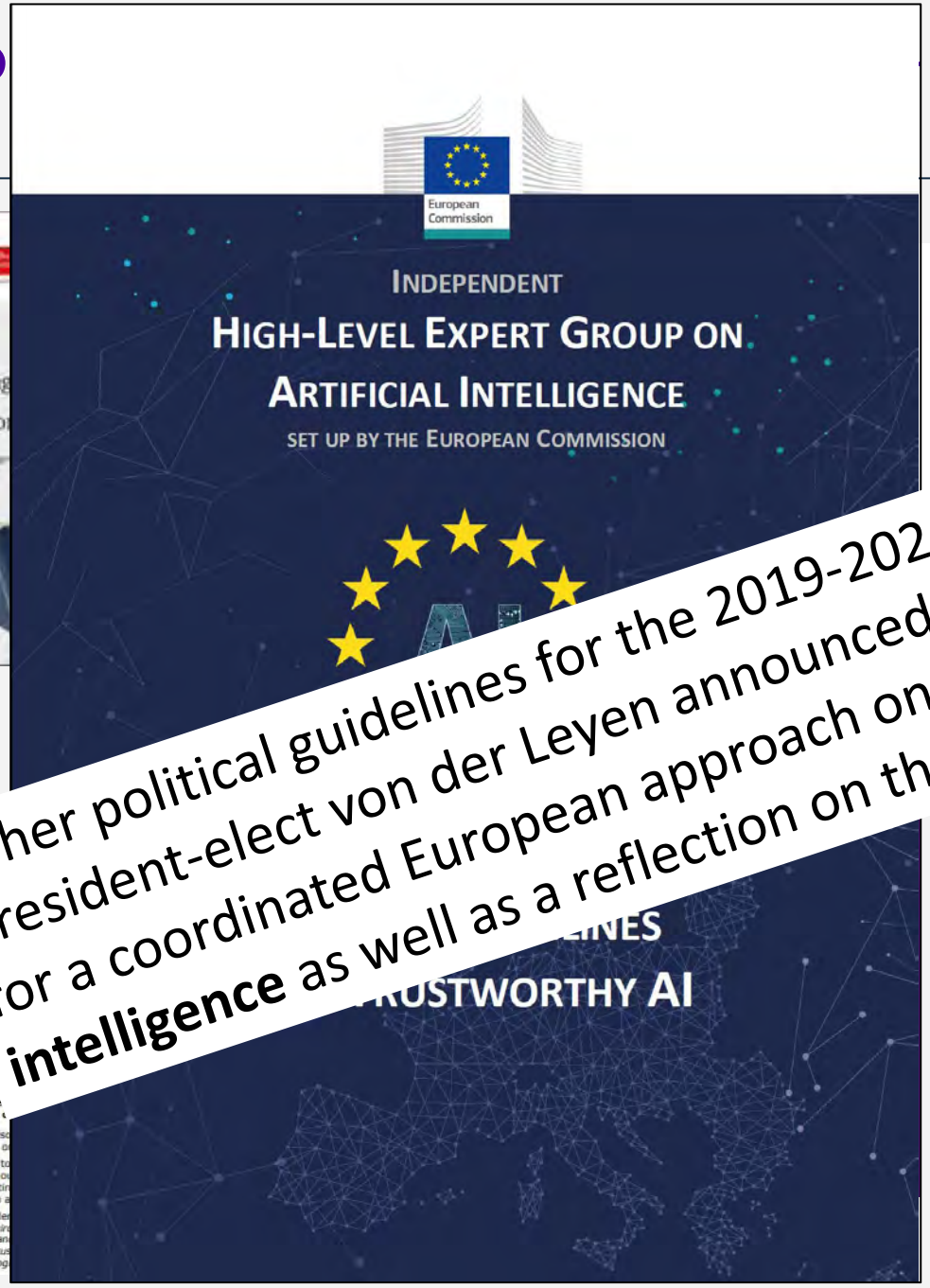


Problem: Complexity of Discourse

P of Discourse

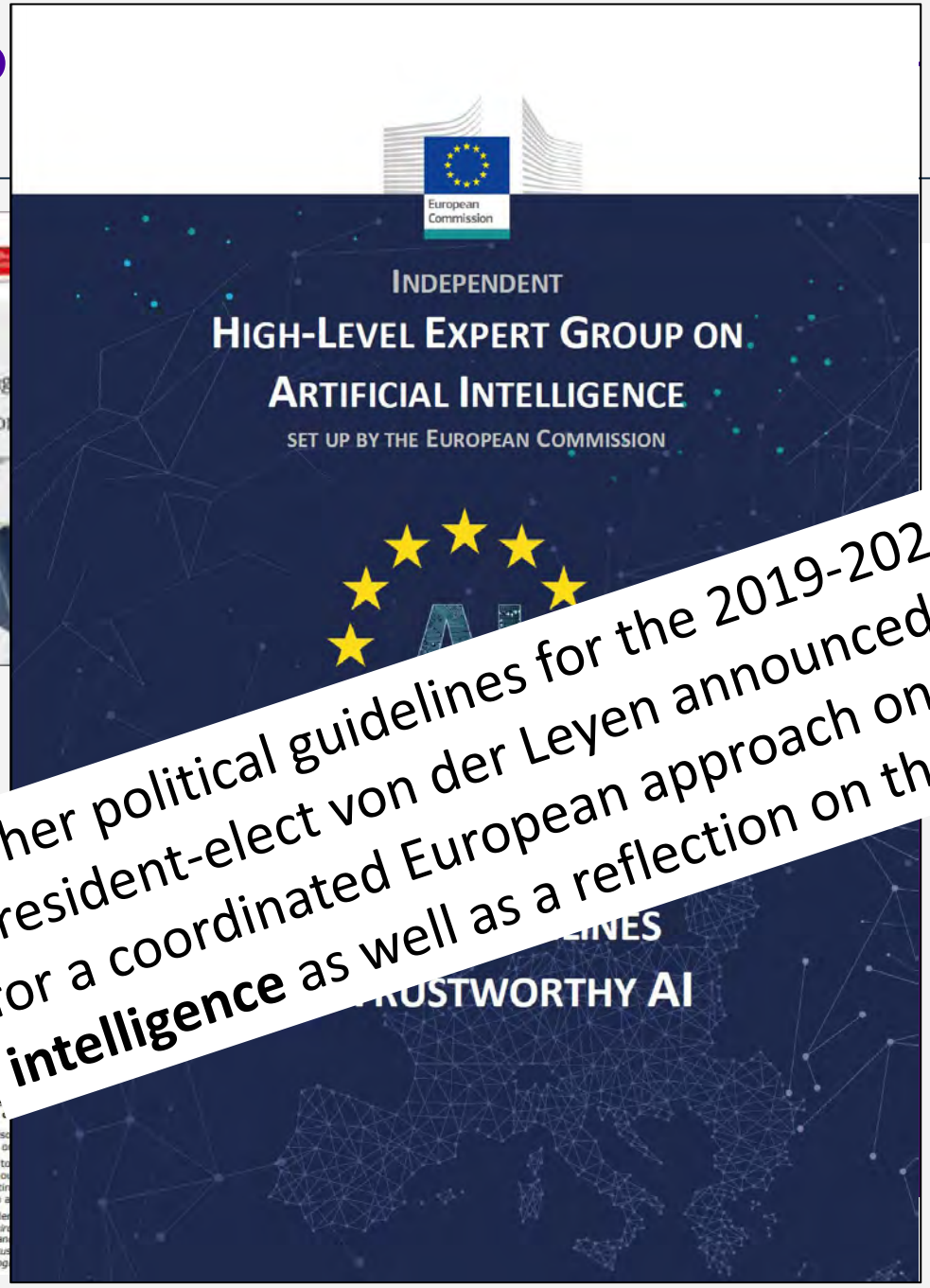


P of Discourse

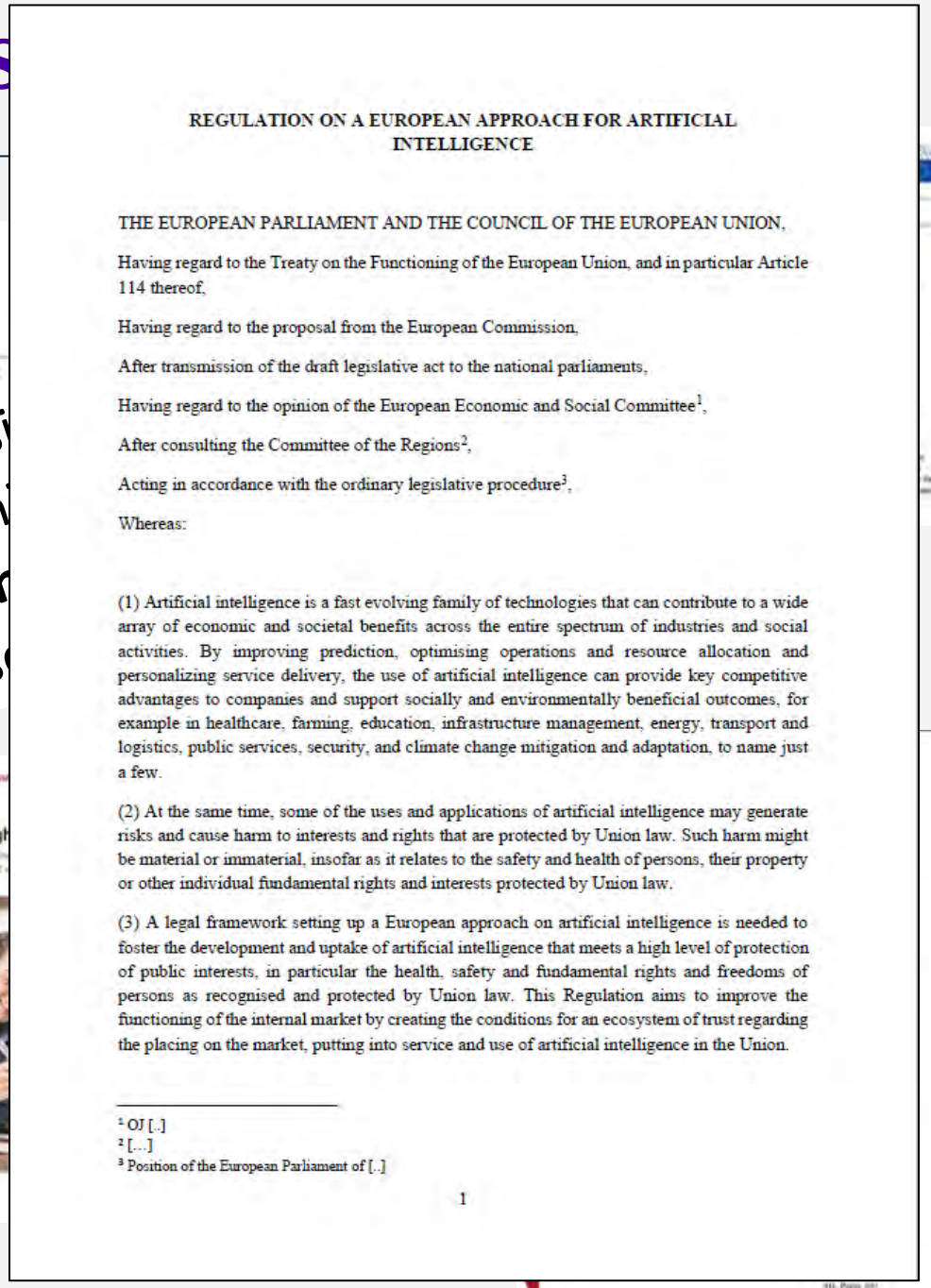


In her political guidelines for the 2019-2024 Commission "A Union that strives for more" President-elect von der Leyen announced the Commission would put forward legislation for a coordinated European approach on the **human and ethical implications of artificial intelligence** as well as a reflection on the better use of big data for innovation.

P of Dis



In her political guidelines for the 2019-2024 Commissi
President-elect von der Leyen announced the Comm
for a coordinated European approach on the **humar**
intelligence as well as a reflection on the better us



REGULATION ON A EUROPEAN APPROACH FOR ARTIFICIAL INTELLIGENCE

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular Article 114 thereof,

Having regard to the proposal from the European Commission,

After transmission of the draft legislative act to the national parliaments,

Having regard to the opinion of the European Economic and Social Committee[1],

After consulting the Committee of the Regions[2],

Acting in accordance with the ordinary legislative procedure[3],

Whereas:

(1) Artificial intelligence is a fast evolving family of technologies that can contribute to a wide array of economic and societal benefits across the entire spectrum of industries and social activities. By improving prediction, optimising operations and resource allocation and personalizing service delivery, the use of artificial intelligence can provide key competitive advantages to companies and support socially and environmentally beneficial outcomes, for example in healthcare, farming, education, infrastructure management, energy, transport and logistics, public services, security, and climate change mitigation and adaptation, to name just a few.

(2) At the same time, some of the uses and applications of artificial intelligence may generate risks and cause harm to interests and rights that are protected by Union law. Such harm might be material or immaterial, insofar as it relates to the safety and health of persons, their property or other individual fundamental rights and interests protected by Union law.

(3) A legal framework setting up a European approach on artificial intelligence is needed to foster the development and uptake of artificial intelligence that meets a high level of protection of public interests, in particular the health, safety and fundamental rights and freedoms of persons as recognised and protected by Union law. This Regulation aims to improve the functioning of the internal market by creating the conditions for an ecosystem of trust regarding the placing on the market, putting into service and use of artificial intelligence in the Union.

[1] OJ [..]
[2] [...]
[3] Position of the European Parliament of [..]

1

# of Dis

COMMITTEE OF MINISTERS
COMITÉ DES MINISTRES

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

MINISTERS' DEPUTIES Recommendations **CM/Rec(2021)8** 3 November 2021

Recommendation CM/Rec(2021)8
of the Committee of Ministers to member States
on the protection of individuals with regard to automatic processing of personal data in the context of profiling

(Adopted by the Committee of Ministers on 3 November 2021
at the 1416th meeting of the Ministers' Deputies)

The Committee of Ministers, under the terms of Article 15.*b* of the Statute of the Council of Europe,

Considering that the aim of the Council of Europe is to achieve a greater unity between its members;

Recalling that digital technologies allow the large-scale processing of data, including personal data, in both the public and private sectors, used for a wide range of purposes including for services widely accepted and valued by society and individuals;

Noting that data are processed in particular by calculation, comparison, correlation and other statistical techniques, with the aim of producing profiles or models that could be used in many ways for different purposes and uses, by matching the data of several individuals;

Considering that, by observing and linking a large amount of data, even anonymous data, profiling techniques can have an impact on the data subjects by placing them in predetermined categories, very often without their knowledge;

Considering that the lack of transparency – or even invisibility – of profiling, and the lack of accuracy that may derive from the automatic application of pre-established rules of inference, can pose significant risks for individuals' rights and freedoms;

Noting that the data processed in the context of profiling may include special categories of personal data, notably biometric data, the misuse of which can cause irreversible damage to data subjects, since such data can be used to access various services and can have legal consequences;

Considering in particular that the protection of fundamental rights, notably the rights to privacy and to protection of personal data, safeguards the existence of different and independent spheres of life where each individual can control his or her information;

Considering the particular vulnerability of some of the persons profiled, including children, and the possible seriousness of the consequences of such profiling, sometimes for the rest of their lives;

Aware of the intensification and diversification of the profiling of individuals, in all spheres of activity;

Website: www.coe.int/cm

House of Commons
Science and Technology Committee

Robotics and artificial intelligence

-2024 Commissi
nced the Commi
ch on the **humar**
on the better us

Council of the European Union

Brussels, 29 November 2021
(OR. en)

14278/21

Interinstitutional File:
2021/0106(COD)

LIMITE

TELECOM 430
JAI 1288
COPEN 412
CYBER 307
DATAPROTECT 267
EJUSTICE 103
COSI 236
IXIM 262
ENFOPOL 465
FREMP 272
RELEX 1012
MI 879
COMPET 860
CODEC 1530

NOTE

From:	Presidency
To:	Delegations
No. Cion doc.:	8115/20
Subject:	Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - Presidency compromise text

I. INTRODUCTION

1. The Commission adopted the proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act, AIA) on 21 April 2021.

COMMITTEE OF MINISTERS
COMITÉ DES MINISTRES

MINISTERS' DEPUTIES Recommendations CM/Rec(2021)8

Recommendation CM/Rec(2021)8 of the Committee of Ministers to member States on the protection of individuals with regard to automatic processing the context of profiling

(Adopted by the Committee of Ministers on 3 November 2021 at the 1416th meeting of the Ministers' Deputies)

The Committee of Ministers, under the terms of Article 15.*b* of the Statute of the Cou

Considering that the aim of the Council of Europe is to achieve a greater unity betwe

Recalling that digital technologies allow the large-scale processing of data, including pers public and private sectors, used for a wide range of purposes including for services widely society and individuals;

Noting that data are processed in particular by calculation, comparison, correlation and c techniques, with the aim of producing profiles or models that could be used in many way and uses, by matching the data of several individuals;

Considering that, by observing and linking a large amount of data, even anonymous techniques can have an impact on the data subjects by placing them in predetermine often without their knowledge;

Considering that the lack of transparency – or even invisibility – of profiling, and the may derive from the automatic application of pre-established rules of inference, can for individuals' rights and freedoms;

Noting that the data processed in the context of profiling may include special categories notably biometric data, the misuse of which can cause irreversible damage to data subj can be used to access various services and can have legal consequences;

Considering in particular that the protection of fundamental rights, notably the rights to p of personal data, safeguards the existence of different and independent spheres of life w control his or her information;

Considering the particular vulnerability of some of the persons profiled, including childre seriousness of the consequences of such profiling, sometimes for the rest of their lives;

Aware of the intensification and diversification of the profiling of individuals, in all sph

Website: www.coe.int/cm

2024/1689 12.7.2024

REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 13 June 2024

laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)

(Text with EEA relevance)

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular Articles 16 and 114 thereof,

Having regard to the proposal from the European Commission,

After transmission of the draft legislative act to the national parliaments,

Having regard to the opinion of the European Economic and Social Committee [1],

Having regard to the opinion of the European Central Bank [2],

Having regard to the opinion of the Committee of the Regions [3],

Acting in accordance with the ordinary legislative procedure [4],

Whereas:

(1) The purpose of this Regulation is to improve the functioning of the internal market by laying down a uniform legal framework in particular for the development, the placing on the market, the putting into service and the use of artificial intelligence systems (AI systems) in the Union, in accordance with Union values, to promote the uptake of human centric and trustworthy artificial intelligence (AI) while ensuring a high level of protection of health, safety, fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (the 'Charter'), including democracy, the rule of law and environmental protection, to protect against the harmful effects of AI systems in the Union, and to support innovation. This Regulation ensures the free movement, cross-border, of AI-based goods and services, thus preventing Member States from imposing restrictions on the development, marketing and use of AI systems, unless explicitly authorised by this Regulation.

(2) This Regulation should be applied in accordance with the values of the Union enshrined as in the Charter, facilitating the protection of natural persons, undertakings, democracy, the rule of law and environmental protection, while boosting innovation and employment and making the Union a leader in the uptake of trustworthy AI.

(3) AI systems can be easily deployed in a large variety of sectors of the economy and many parts of society, including across borders, and can easily circulate throughout the Union. Certain Member States have already explored the adoption of national rules to ensure that AI is trustworthy and safe and is developed and used in accordance with fundamental rights obligations. Diverging national rules may lead to the fragmentation of the internal market and may decrease legal certainty for operators that develop, import or use AI systems. A consistent and high level of protection throughout the Union should therefore be ensured in order to achieve trustworthy AI, while divergences hampering the free circulation, innovation, deployment and the uptake of AI systems and related products and services within the internal market should be prevented by laying down uniform obligations for operators and

[1] OJ C 517, 22.12.2021, p. 56.
[2] OJ C 115, 11.3.2022, p. 5.
[3] OJ C 97, 28.2.2022, p. 60.
[4] Position of the European Parliament of 13 March 2024 (not yet published in the Official Journal) and decision of the Council of 21 May 2024.

Brussels, 29 November 2021
(OR. en)

14278/21

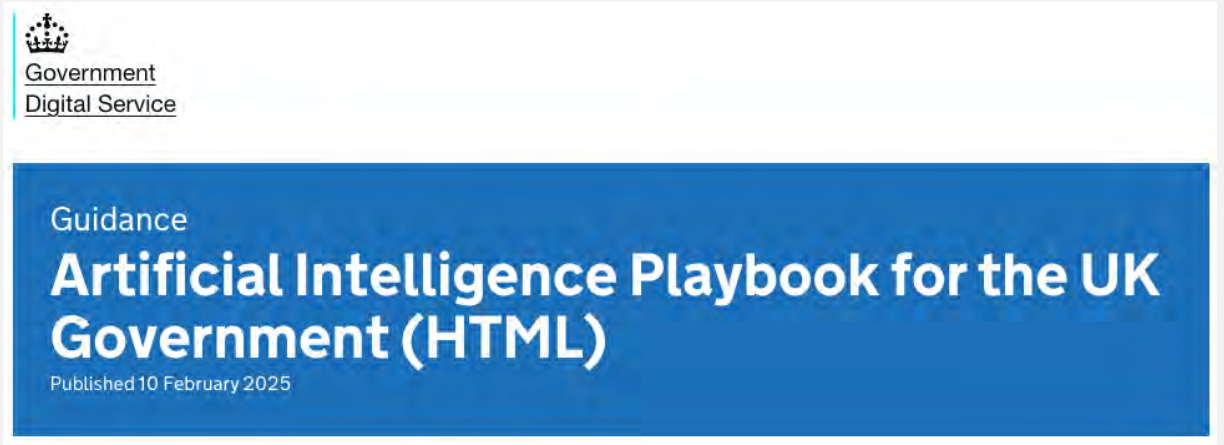
LIMITE

TELECOM 430
JAI 1288
COPEN 412
CYBER 307
DATAPROTECT 267
EJUSTICE 103
COSI 236
IXIM 262
ENFOPOL 465
FREMP 272
RELEX 1012
MI 879
COMPET 860
CODEC 1530

a Regulation of the European Parliament and of the Council
harmonised rules on artificial intelligence (Artificial Intelligence
ending certain Union legislative acts
compromise text

d the proposal for a Regulation laying down harmonised rules
Artificial Intelligence Act, AIA) on 21 April 2021.

In the last few weeks



In

SERVICE DE PRESSE

Palais de l'Élysée, Tuesday February 11th 2025

AI ACTION SUMMIT

Co-chaired by France and India

10-11 February, 2025, Paris

Statement[1] on Inclusive and Sustainable Artificial Intelligence for People and the Planet

1. **Participants from over 100 countries, including government leaders, international organisations, representatives of civil society, the private sector, and the academic and research communities gathered in Paris on 10 and 11 February 2025 to hold the AI Action Summit.** Rapid development of AI technologies represents a major paradigm shift, impacting our citizens, and societies in many ways. In line with the Paris Pact for People and the Planet, and the principles that countries must have ownership of their transition strategies, we have identified priorities and launched concrete actions to advance the public interest and to bridge digital divides through accelerating progress towards the SDGs. Our actions are grounded in three main principles of science, solutions - focusing on open AI models in compliance with countries frameworks - and policy standards, in line with international frameworks.

2. **This Summit has highlighted the importance of reinforcing the diversity of the AI ecosystem.** It has laid an open, multi-stakeholder and inclusive approach that will enable AI to be human rights based, human-centric, ethical, safe, secure and trustworthy while also stressing the need and urgency to narrow the inequalities and assist developing countries in artificial intelligence capacity-building so they can build AI capacities.

3. Acknowledging existing multilateral initiatives on AI, including the United Nations General Assembly Resolutions, the Global Digital Compact, the UNESCO Recommendation on Ethics of AI, the African Union Continental AI Strategy, and the works of the Organization for Economic Cooperation and Development (OECD), the council of Europe and European Union, the G7 including the Hiroshima AI Process and G20, **we have affirmed the following main priorities:**

   - Promoting AI accessibility to reduce digital divides;
   - Ensuring AI is open, inclusive, transparent, ethical, safe, secure and trustworthy, taking into account international frameworks for all
   - Making innovation in AI thrive by enabling conditions for its development and avoiding market concentration driving industrial recovery and development
   - Encouraging AI deployment that positively shapes the future of work and labour markets and delivers opportunity for sustainable growth
   - Making AI sustainable for people and the planet
   - Reinforcing international cooperation to promote coordination in international governance

To deliver on these priorities:

- [2]**Founding members have launched a major Public Interest AI Platform and Incubator**, to support, amplify, decrease fragmentation between existing public and private initiatives on Public Interest AI and address digital divides. The Public interest AI Initiative will sustain and

[1] In line with the approach of previous Summits, this Statement relates to civil applications and use of AI only
[2] Kenya, Germany, Chile, Finland, Slovenia, France, Nigeria, Morocco, India

Reuters

World Business Markets Sustainability Legal Breakingviews Technology Investigat

Vance tells Europeans that heavy regulation could kill AI

By **Jeffrey Dastin** and **Ingrid Melander**

February 11, 2025 4:41 PM GMT · Updated 15 days ago

Summary Companies

- Vance says Europeans risk killing AI with their red tape
- US, UK do not sign summit communique
- Vance says Trump will ensure US remains lead AI player

SHERPA PROJECT

The Ethical and Human Rights Implications of AI

www.project-sherpa.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

What was SHERPA?

- EU funded project, with 11 partners
- 3½ years - finished in October 2021
- What we did (all available online):
  - 10 case studies
  - 5 scenarios
  - Focus groups
  - Interviews with experts and stakeholders.
  - Online survey
  - Delphi study
  - A series of briefings
  - Interviews with MEPs - advocacy

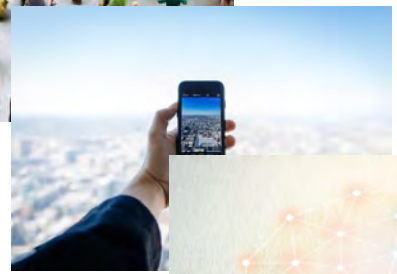
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

10 Case Studies

SHERPA

IoT

Government



Sustainability - Smart Cities

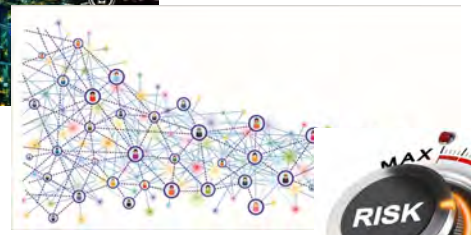
Science

Agriculture

Insurance

Energy and Utilities

Retail and Trade



Manufacturing

Communication, Media

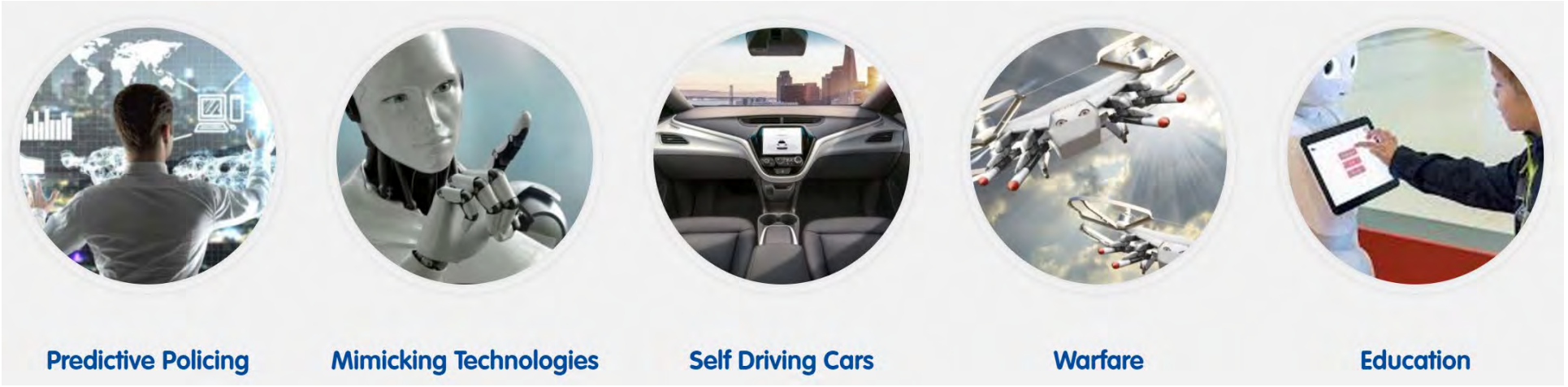
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

DE MONTFORT UNIVERSITY LEICESTER

SHERPA

Scenarios

https://www.project-sherpa.eu/scenarios/



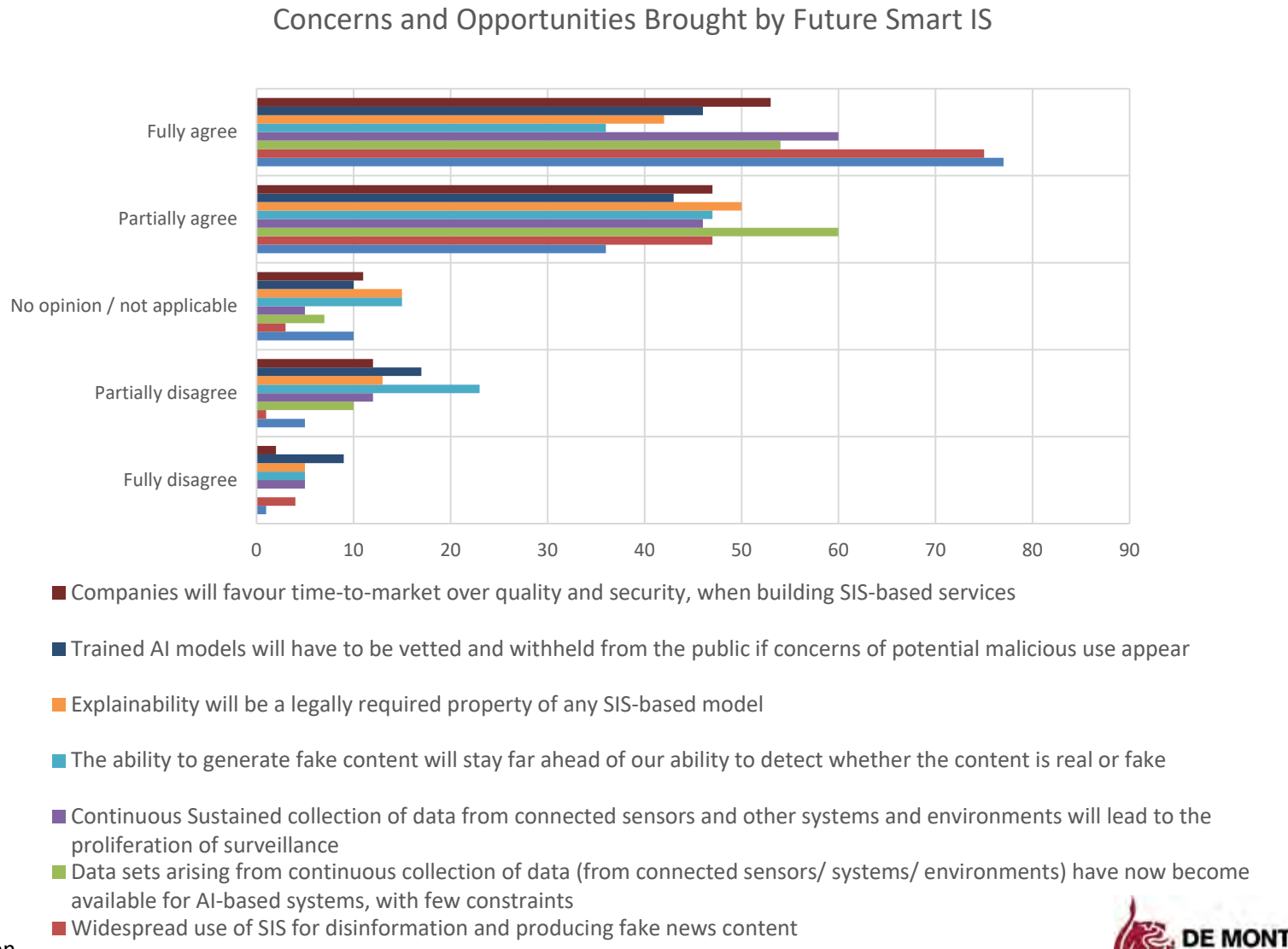
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

DE MONTFORT UNIVERSITY LEICESTER

Survey Results

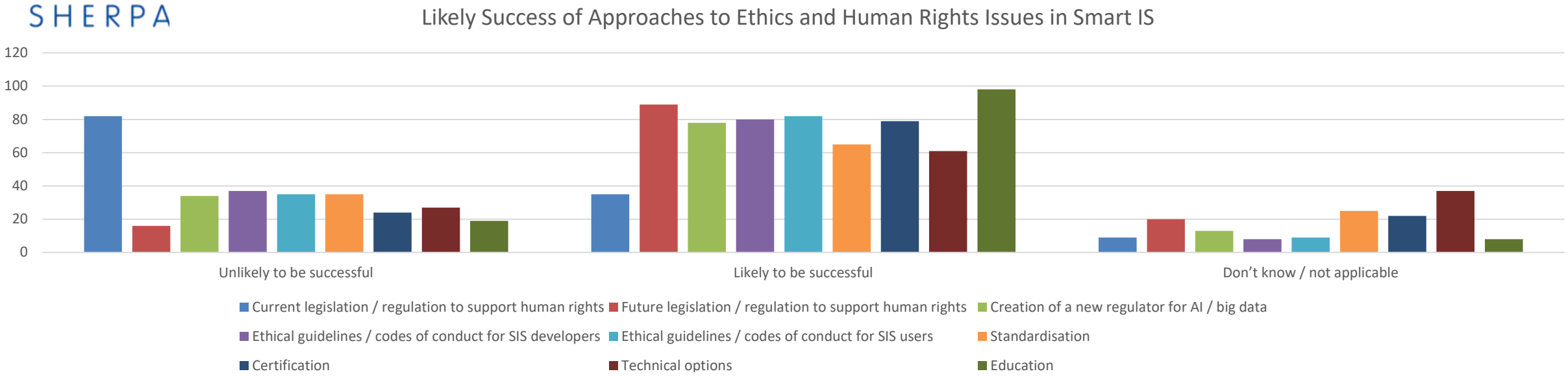
Around 140 respondents

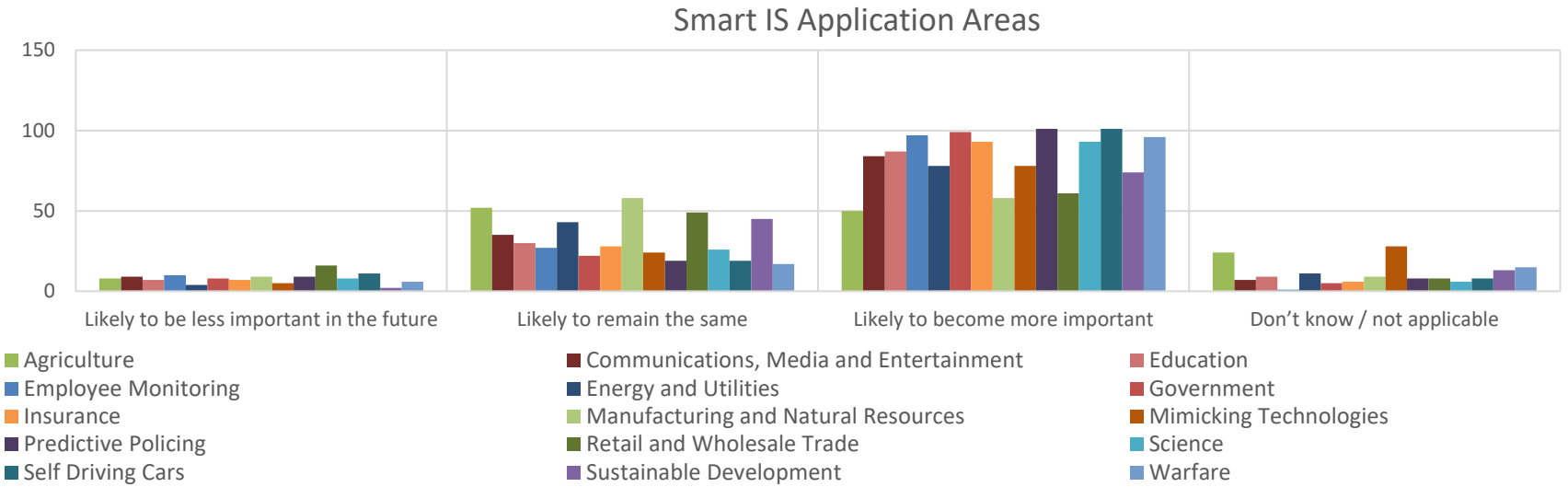
- Roughly equal gender balance
- Mainly European, but some more global
- Age 23-80, most between 41-60
- Self reported good level of knowledge of Smart IS
- Full report available as SHERPA deliverable 2.3



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

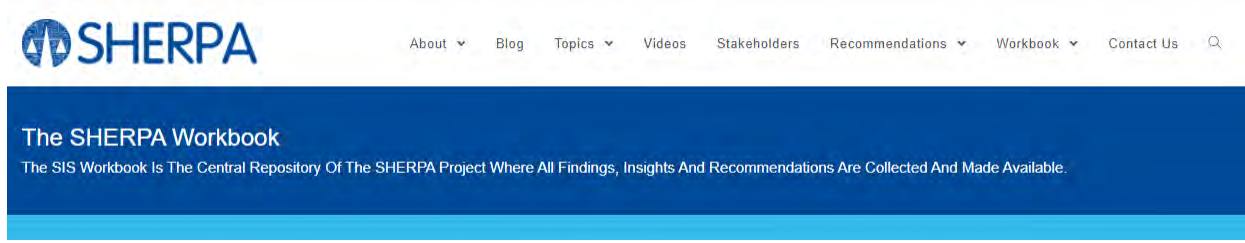
Survey Results

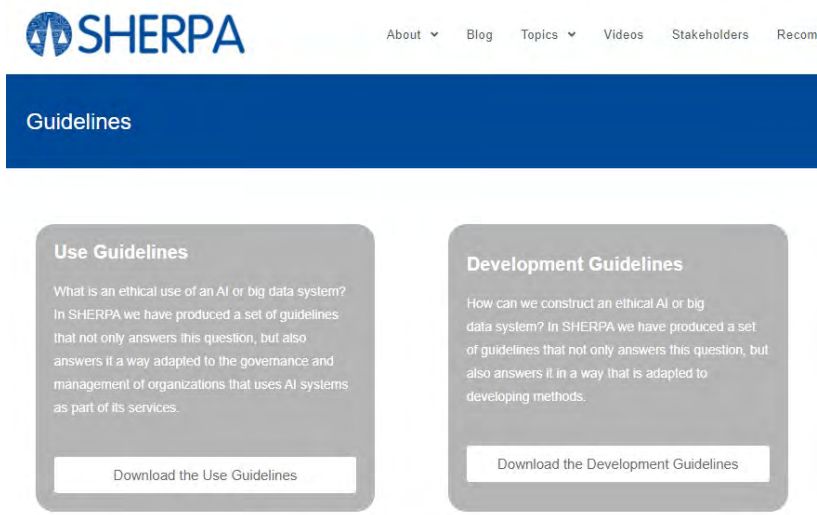




SHERPA – More Results

- Outcomes:
  - A workbook (online)
  - A set of recommendations (online)
  - Regulation
  - Standardisation
  - Guidelines
    - For developers & for users
  - Ethics by design – taken up by Horizon Europe as part of their approach to research



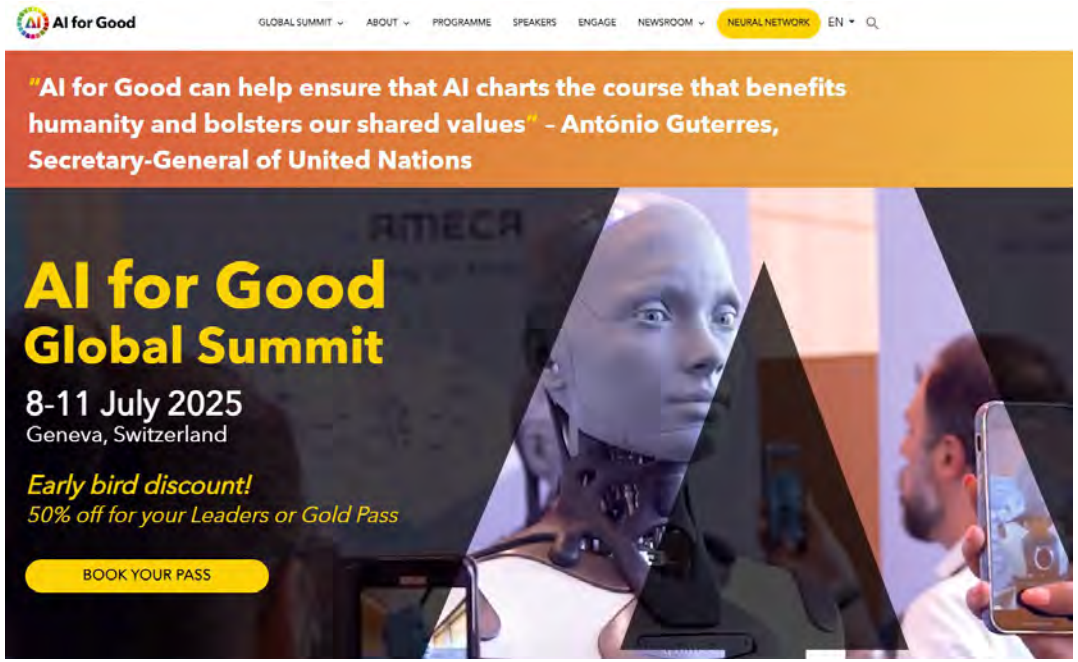


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

SHERPA

Desired outcomes

- Economic growth for all
- Addressing global challenges (Sustainable Development Goals) & societal missions
- Better (personalised) services
- Increased human capabilities (compensate disabilities)
- Inclusion & democratic participation
- Empowerment



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

DE MONTFORT UNIVERSITY LEICESTER

SHERPA

AI Stakeholders

Policy

- EU
- Funders
- National Policy
- Regulators
- Ethics Bodies
- International Bodies

Individuals

- Users
- Activists
- Lay Experts
- Developers
- Non-users

Organisations

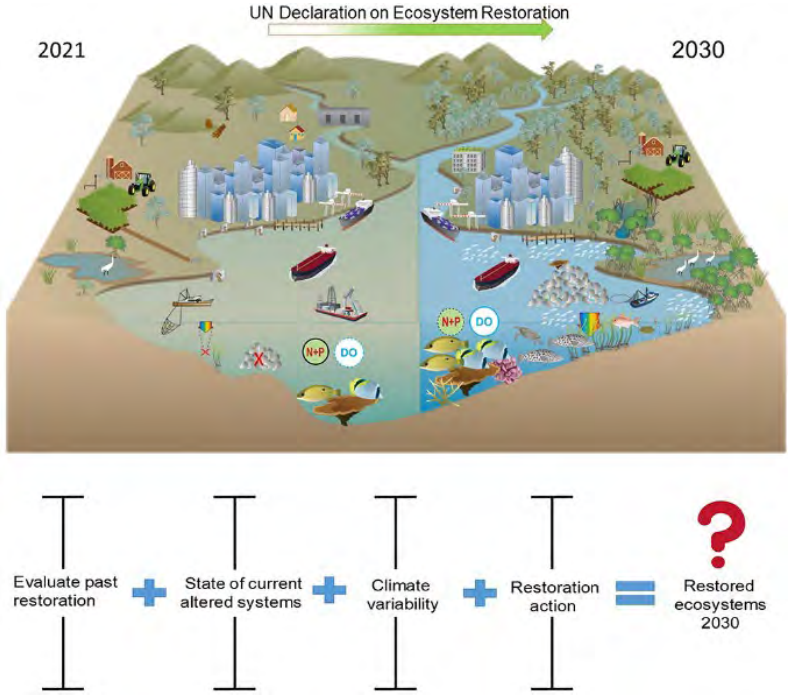
- Developers
- Deployers
- Users
- Professional Bodies
- Media
- Educators
- Standards Bodies

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

DE MONTFORT UNIVERSITY LEICESTER

Insights

- There is no magic bullet
  - epistemic complexity
  - distribution of responsibility
  - technical progress
- What is required is an intelligent mix of options
- Ecosystem metaphor



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

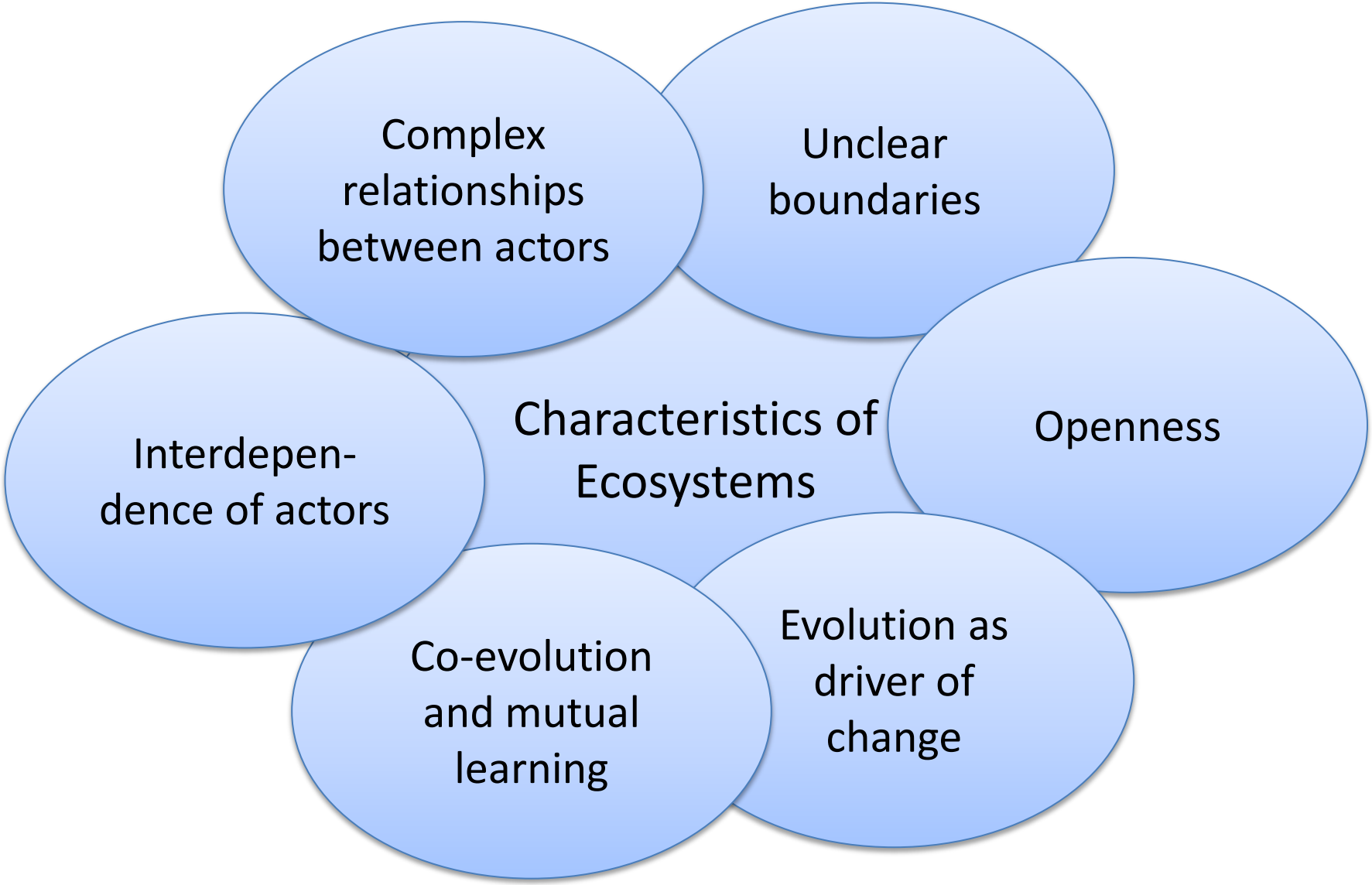
SHERPA

Towards an Ecosystem of Artificial Intelligence for Human Flourishing

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

DE MONTFORT UNIVERSITY LEICESTER

SHERPA

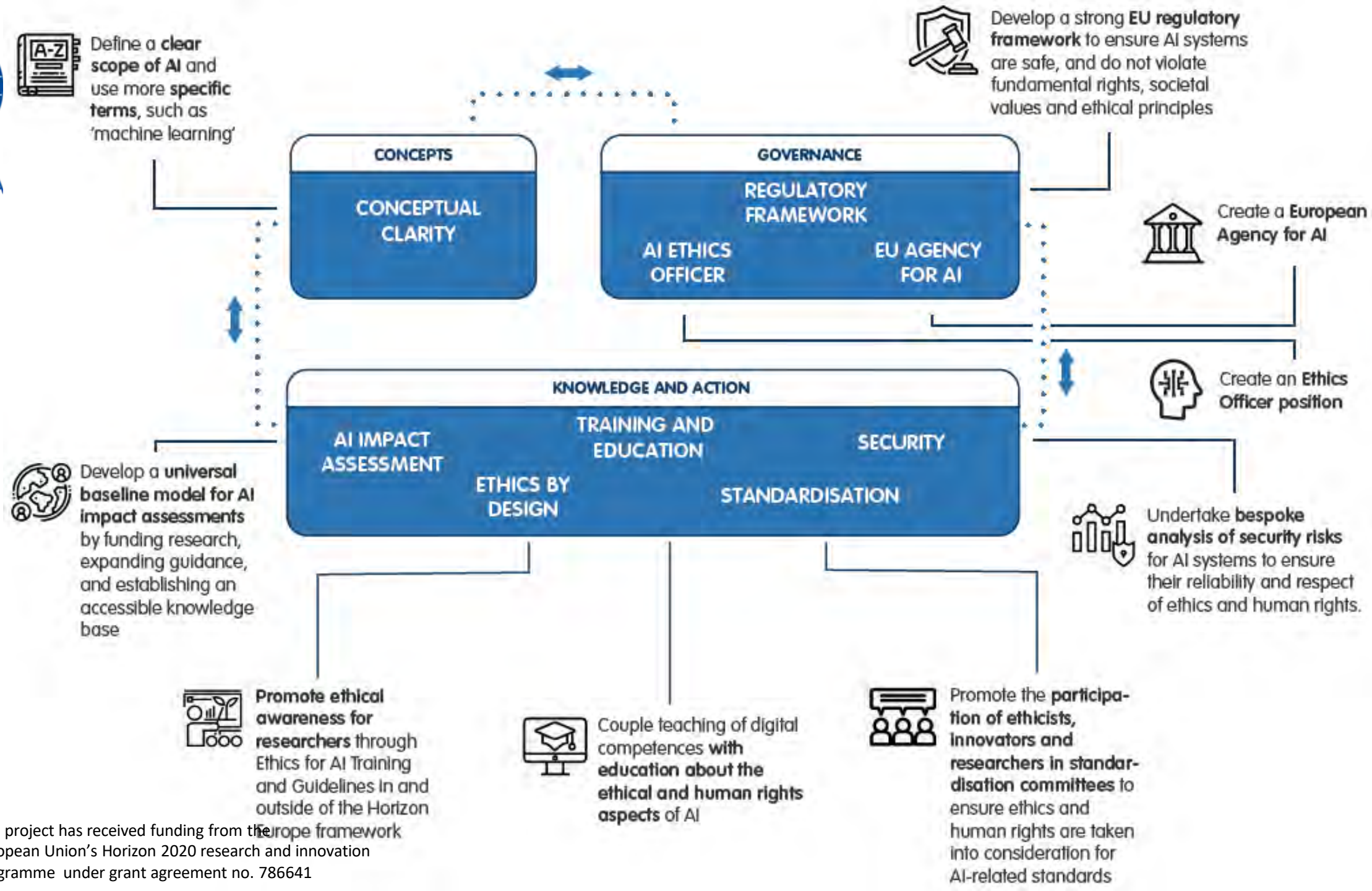


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

DE MONTFORT UNIVERSITY LEICESTER

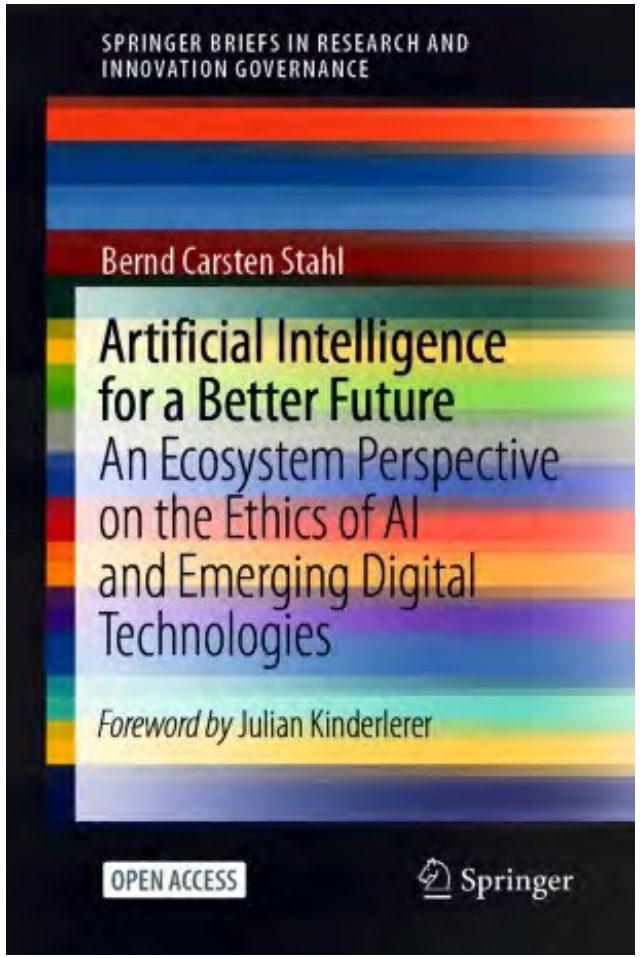
SHERPA

Recommendations Overview



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

Sherpa Based Open Access Book



Artificial Intelligence for a Better Future

by Bernd Stahl

https://link.springer.com/book/10.1007%2F978-3-030-69978-9

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 786641

SHERPA

Recommendations Overview

After SHERPA - What next?

Define a **clear scope of AI** and

Develop a strong **EU regulatory framework** to ensure AI systems are safe, and do not violate rights, societal thical principles

Create a **European Agency for AI**

Create an **Ethics Officer position**

D
b
ir
b
e
a
a
b

Undertake **bespoke analysis of security risks** for AI systems to ensure their reliability and respect of ethics and human rights.

ipa-

ndar-
es to

outside of the Horizon Europe framework

aspects of AI

ensure ethics and human rights are taken into consideration for AI-related standards

TECHETHOS
FUTURE ○ TECHNOLOGY ○ ETHICS

https://www.techethos.eu/

TechEthos – Ethics of Emerging Technologies

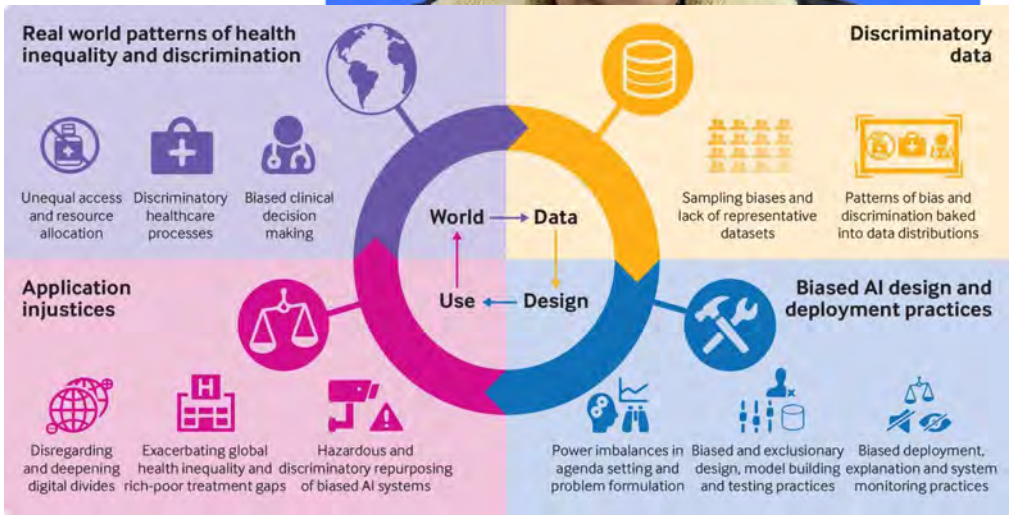
TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

Introduction

The central problem for the ethics of emerging technologies is that we humans cannot predict the future, and therefore do not know which ethical issues will play out once the technology is fully developed and entrenched in society. As the emerging technology is still evolving, many questions can arise about its nature, its future use, and its social consequences.

In some ways the future is already here…

- Need for some way to analyse ethical issues in emerging technologies?



Inequality and discrimination in the design and use of AI in healthcare applications, Source: British Medical Journal

TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

NEWSLETTERS
Sign up to read our regular email newsletters

NewScientist

News Podcasts Video Technology Space Physics Health More Shop Courses Events

Drones may have attacked humans fully autonomously for the first time

TECHNOLOGY 27 May 2021

By David Hambling

The Kargu-2 quadcopter is armed with an explosive charge and can attack autonomously
EMRE CAVDAR/STM

Military drones may have autonomously attacked humans for the first time ever last year, according to a United Nations report. While the full details of the incident, which took place in Libya, haven't been released and it is unclear if there were any casualties, the event suggests that international efforts to ban lethal autonomous weapons before they are used may already be too late.

Met Police: Live facial recognition cameras result in 17 arrests in south London

25 March

PA MEDIA

The technology has been used multiple times in Croydon, south London

By Jess Warren
BBC News

What Are Emerging Technologies?

Five key attributes that appear to help identify a technology as emerging (Rotolo et al., 2015):

- radical novelty,
- relatively fast growth,
- coherence (persisting over time),
- prominent impact (on the socio-economic domain), and
- uncertainty and ambiguity (as we don't really know what the future holds and therefore what the impact of a technology will be).

TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

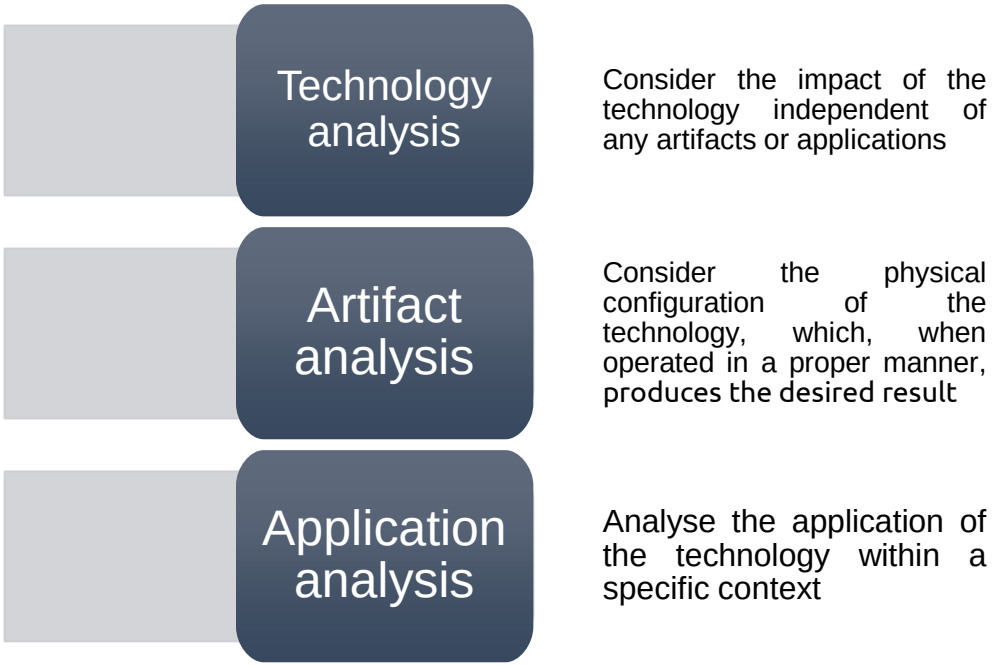
Existing Ethical Frameworks I

Three previous approaches to ethical analysis:

- Anticipatory Technology Ethics (ATE) working towards ATE+
- Ethical Impact Assessment (EIA) and
- Future Studies

Anticipatory Technology Ethics (ATE) (Brey 2012) & ATE+ (Umbrello et al., 2023)

- This approach has 3 levels & focuses on emerging technologies from the perspective of trying to identify what is both good and bad about them
- Critique - trying to predict what might be the impact and outcomes of emerging technologies - difficult to recognise what might be the unintended and emergent properties.
- An expanded version of ATE, named ATE+, has been developed:
  - evaluation of 'what is good' is related to practice
  - reflections on 'whose values'
  - evaluation requires engineering & user expertise, as well as context
  - focuses also on ethical opportunities not just challenges
- This augmented version aims to be more useful in applied settings, in particular complementing ethics-by-design approaches.



TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

Existing Ethical Frameworks II and III

2. Ethical Impact Assessment (EIA) (Wright and Friedewald, 2013)

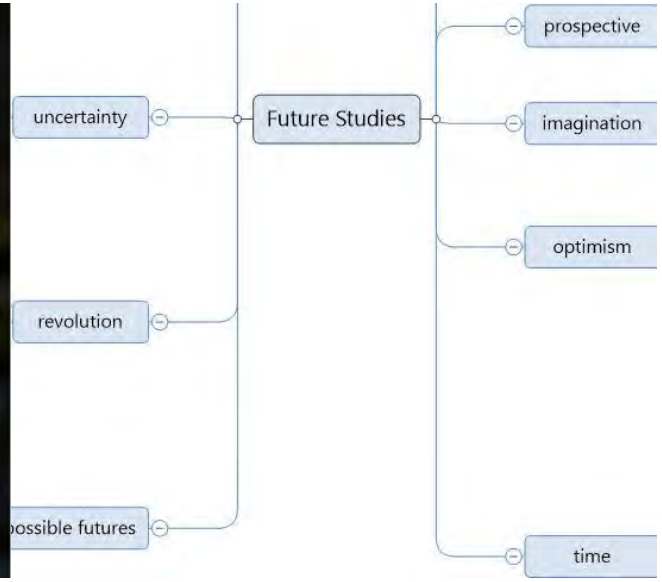
- The aim of this framework is to facilitate consideration of ethical issues, in consultation with stakeholders, which may arise in their undertaking but does not account for emerging technologies.

The EIA framework consists of the following steps:

1) conducting an EIA threshold analysis,
2) preparing an EIA plan,
3) identifying ethical impacts
4) evaluating the ethical impacts (step 3 and 4 are to be carried out in consultation with stakeholders),
5) formulating and implementing remedial actions,
6) reviewing and auditing the EIA.

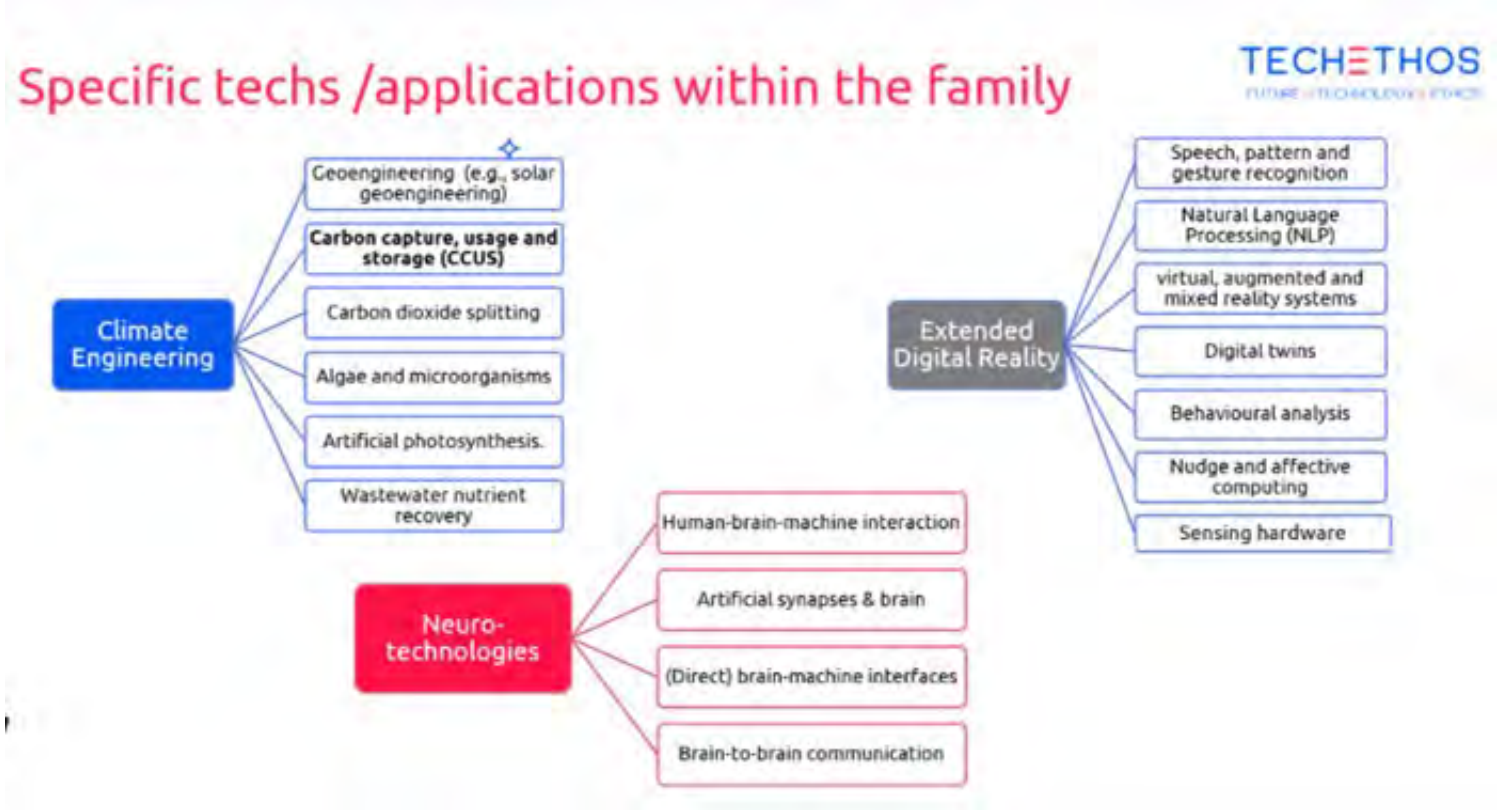
3. Future Studies (e.g. Sardar 2010)

- Future Studies emerges as an interdisciplinary field, recognising that the 'future' is not produced by one agent, but a number of intersecting, often colliding and reacting processes, which is often also seen as technologies emerge.



TechEthos

- H2020 3 year funded project to look at the ethics of emerging technologies.
- Selected 3 technology families to focus on: Climate Engineering, Neurotechnologies and digital Extended Reality (dXR).



TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

The Challenge

How can we prioritise ethics and societal values in the design, development and deployment of new and emerging technologies, particularly those with high socio-economic impact?

New and emerging technologies are expected to generate new opportunities and offer a wealth of socio-economic benefits. However, in the early stages of their development, these technologies also pose a number of potential ethical challenges and societal consequences.

TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

The Vision

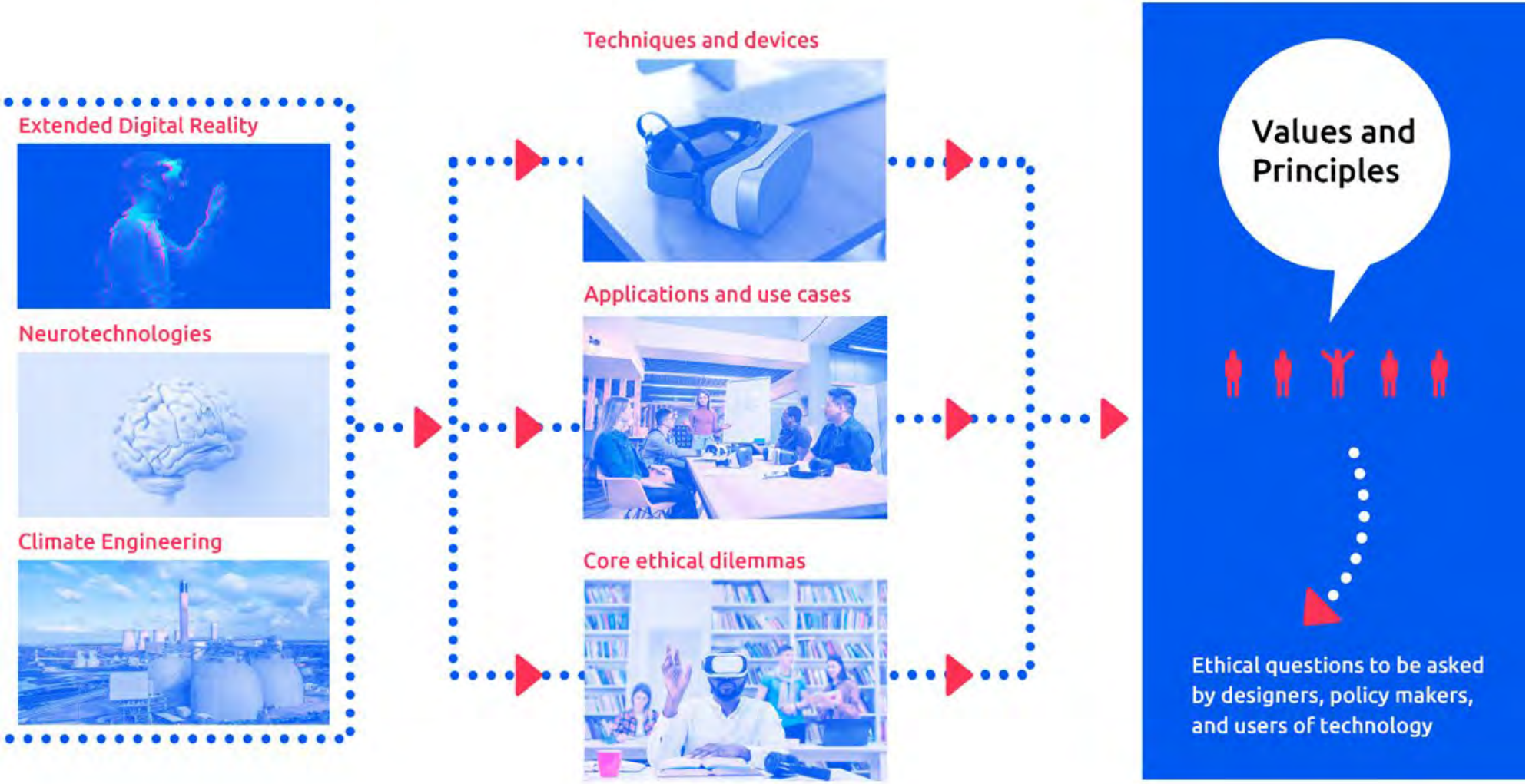
Ethics by design, or in other words, bringing ethical and societal values into the design and development of technology from early on in the design and development process.

With this principle in mind, TechEthos aimed to produce and ethics framework and ethics guidelines for three technologies, ensuring that they work for different actors in the field such as researchers, research ethics committees and policy makers.

To reconcile the needs of research and innovation and the concerns of society, TechEthos has explored the awareness, acceptance and aspirations of academia, industry and the general public alike and will reflect them in the guidelines.

TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

Three Roads to Arrive at Values and Principles



Cross Cutting Topics in Ethical Issues

Narratives of lay ethics	•What is the general perception the ethical issues with new technologies, e.g. opening of Pandora's box, messing with nature, be careful what you wish for
Irreversibility	•Whether what can be done can be reversed, or are there points at which we have travelled too far along a specific path, e.g. climate change initiative making the situation worse, the changes in society arising from the blending of physical and virtual worlds
Novelty and speed of change	•Are changes happening too fast, over-inflated expectations, novelty and uncertainty, fear of missing out (fomo)
Vulnerability and the structures of power	•Concerned with distributive justice, inexorably, as a result of political, cultural, economic contexts, need to situate in 'real world' of inequalities and injustices structures, and ask questions about 'who' is influencing and being most affected.
Governance of uncertainty	•What these technologies mean for us now and how we might control them, such as by regulation.
Perception of uncertainty	•We cannot (yet) foresee the future, but we do try and think about what the future possibilities/scenarios might look like and their implications, can we use these technologies to 'rewrite the future' and how much control might we have over that
Security	•An essential ethical concept because it is necessary to preserve the ethical design of any application but also balance/maintain with security
Ethics washing	•Learnt from the AI field, ie., pushing for an ethical governance of AI in order to avoid hard laws that could limit technological innovations.

TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

Values and principles in NLP Questions

Natural Language Processing (NLP) I

Autonomy	Can one limit moral projections onto chatbots?
Dignity	Can conversation data be used to imitate someone's speech in ways that threaten or challenge their dignity?
Decency	How to make sure that chatbots do not insult or demean human subjects? How should chatbots respond to insults?
Non-manipulation	How to deal with chatbots designed for nudging or eliciting a particular response?
Respect of cultural differences	How can chatbots be adapted for a particular audience, culture, or dialect?

Natural Language Processing (NLP) II

Avoiding Bias	How can a chatbot address a human without prejudice for gender, race, sexuality, etc.?
Responsibility	Who should be responsible for chatbot malfunctioning?
Privacy	When can a chatbot disclose a private conversation?
Security and Traceability	How to make sure that the chatbot remains secure against manipulation?

TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

Values and principles in XR Questions

eXtended Reality (XR) I

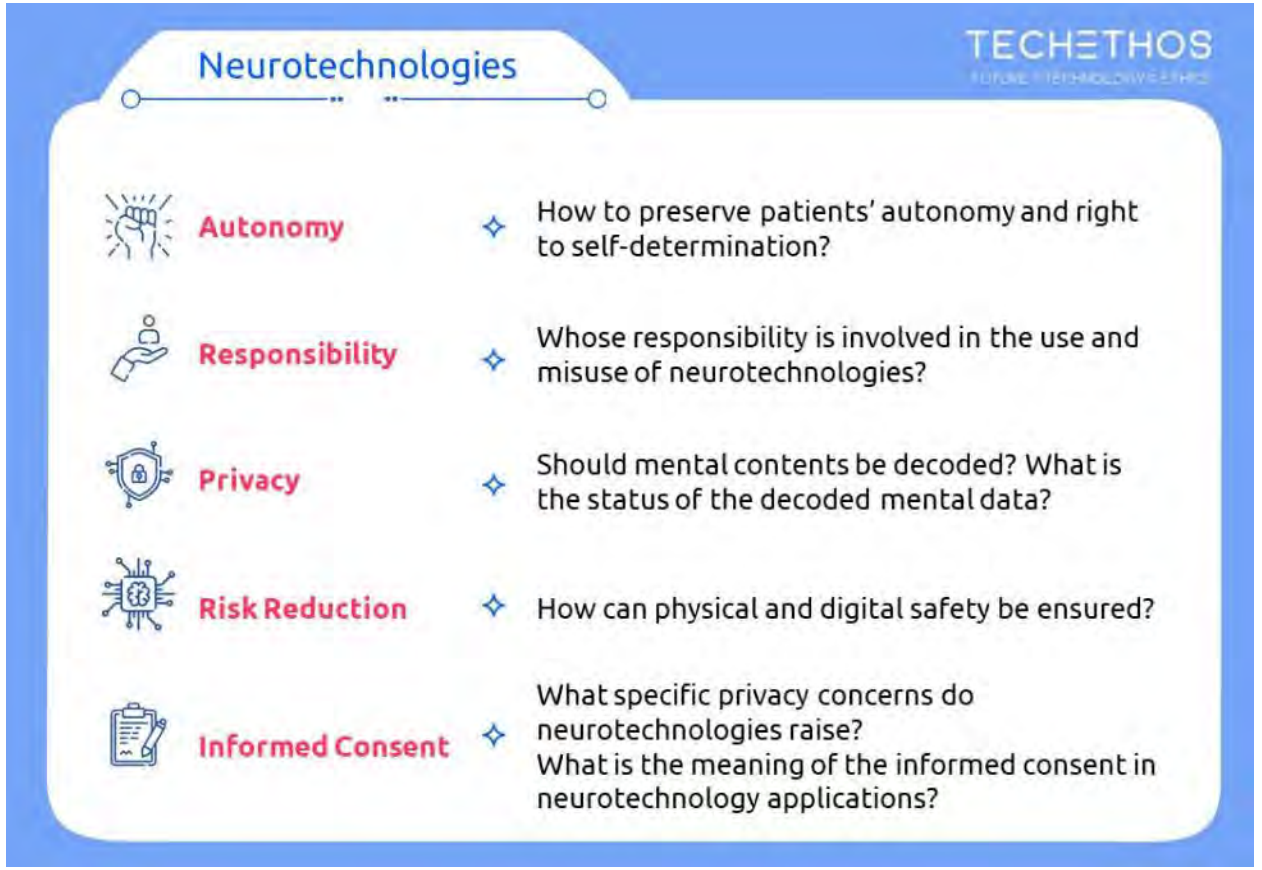
Transparency	Should there be limits for immersion?
Dignity	Can avatars simulate the presence of individuals, including the dead?
Privacy	How to address privacy concerns raised by XR?
Non-manipulation	Can nudging be controlled in XR?
Responsibility	Should real-world sanctions be issued for virtual misconduct?

eXtended Reality (XR) II

Environmental and security risk reduction	How can physical and digital safety be ensured in XR applications?
Dual use and misuse	Can XR be exploited for malicious purposes?
Power	How can social justice be respected in a metaverse and its material implications?
Labour	How can just labour and economic conditions be ensured in the metaverse?
Bias	How will XR representations influence gender issues?

TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

Values and principles in Neurotechnologies Questions



TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

Values and principles in Climate Engineering Questions

Carbon Dioxide Removal (CDR)

Distributive justice	How can costs of climate engineering be distributed in a just way?
Procedural justice	How to include all affected parties in the decision making?
Future responsibility	How to act responsibly in view of future generations?
Side-effects	Are side-effects of climate engineering worse than their climate benefits?

Solar Radiation Management (SRM)

Distributive justice	How can costs of climate engineering be distributed in a just way?
Procedural justice	How to include all affected parties in the decision making?
SRM research ethics	Does research make implementation more likely?
SRM termination shock	Can the termination be catastrophic?

TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

Approach I

The approach integrated the theoretical ethical frameworks with two types of 'hands-on' information: 1) policy documents, and 2) empirical data concerning ethical issues of the technologies, as drawn from industry and academic experts

- Integrating ethics with policy - scan of existing ethical frameworks, 20 per tech family
- Map the characteristics of the extracted frameworks to make sure there was a sufficiently diverse variety of policy documents - particularly to ensure that a mix of academic as well as grey literature articles had been captured. Such as:

Guideline	Type of organisation	Definition	Extract of source guideline
Ethical code	Academia	Ethical codes set forth responsibilities to which individuals and groups or organisations hold themselves to account.	...professional self-regulation […] should start within a company, institution or other work unit with a code of ethics or set of clearly articulated principles to which leadership adheres... (Chang et al 2019)

TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

Approach II

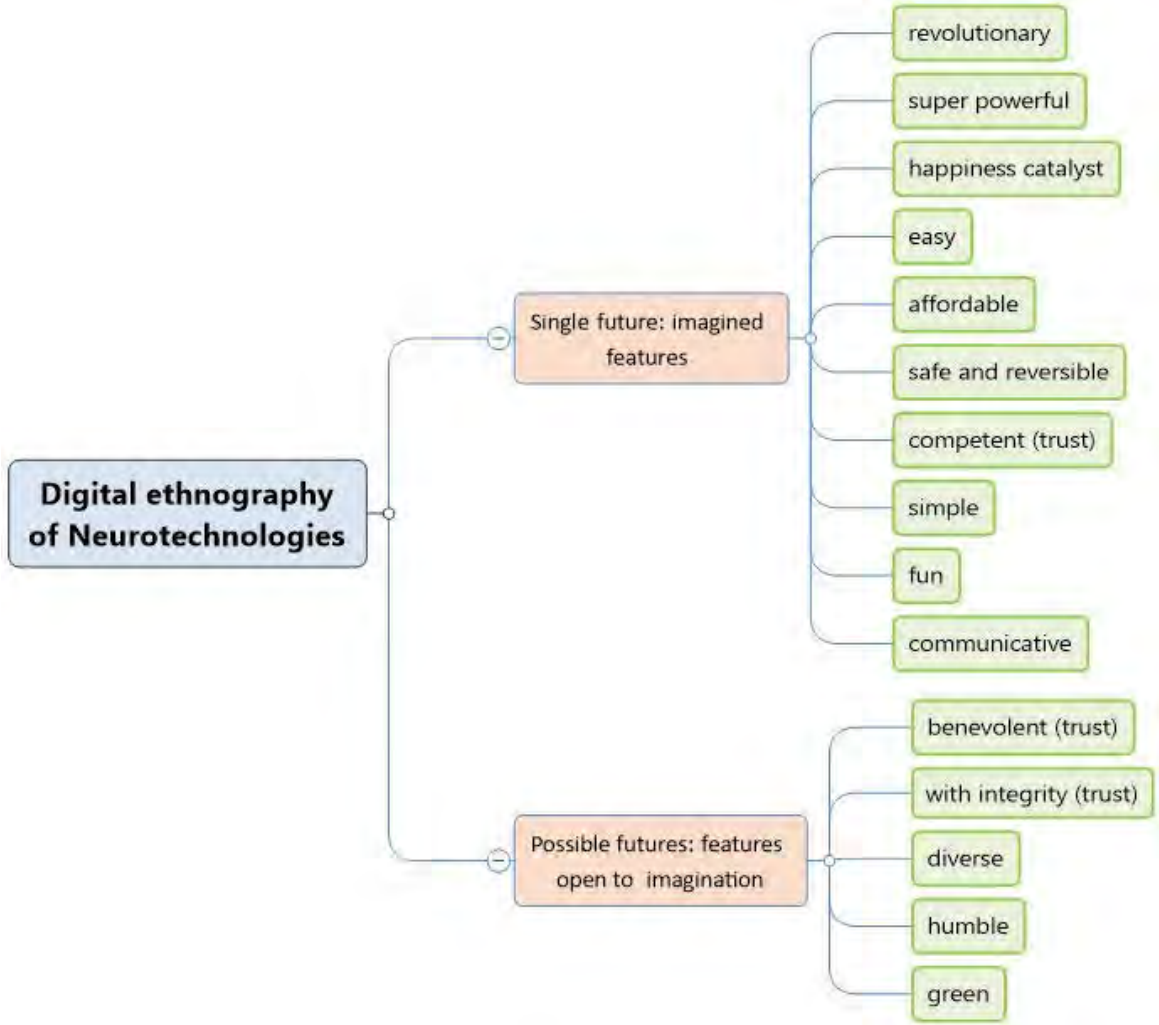
TECHETHOS
FUTURE ○ TECHNOLOGY ○ ETHICS

Following the mapping of policy documents, the ethical frameworks could be integrated with primary data.

This involved identifying & analysing imaginaries of the future envisioned by tech-developers (see example), in order to speculate on future ethical issues that the technology families might bring.

Digital ethnographies:

- A classic definition of traditional ethnography: "describe the lives of people other than ourselves, with an accuracy and sensitivity honed by detailed observation and prolonged first-hand experience" (Ingold cited in Pink & Morgan, 2013).
- Covid-19 restrictions meant we had to turn to 'digital ethnography'…
- A search for businesses' proposing applications within the technology families has been made from the business platform LinkedIn.
- The sample included website pages and YouTube videos (12 in total).



Example: identifying imaginaries of the future envisioned by developers of Neurotechnology

TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

Approach III

Experts' consultations: interviews and workshop

- During the expert interviews, ethical dilemmas, questions informed by epistemological analysis, future studies, as well as the 'guiding questions' method suggested by Stahl, Timmermans and Flick (2017) have been used in order to open ethical reflection on new and emerging issues.
- 8 interviews with experts, across the range of tech families and a range of countries globally, lasting approximately 30 minutes.
- Consultation with 10 European ethics experts through an online workshop, June 2022
- The consultation with experts was conducted through qualitative interviews and workshops that were set up to receive feedback on the following questions:
  - Clarity: Is the meaning of the value in the context of this technology family clear and comprehensible?
  - Completeness: Is the main argument in the subsection complete? What should be added?
  - Operationalization: Are the questions at the end of the subsection helpful operationally? Is anything missing in that aspect?
  - What else do you find interesting and worth mentioning about this technology family?

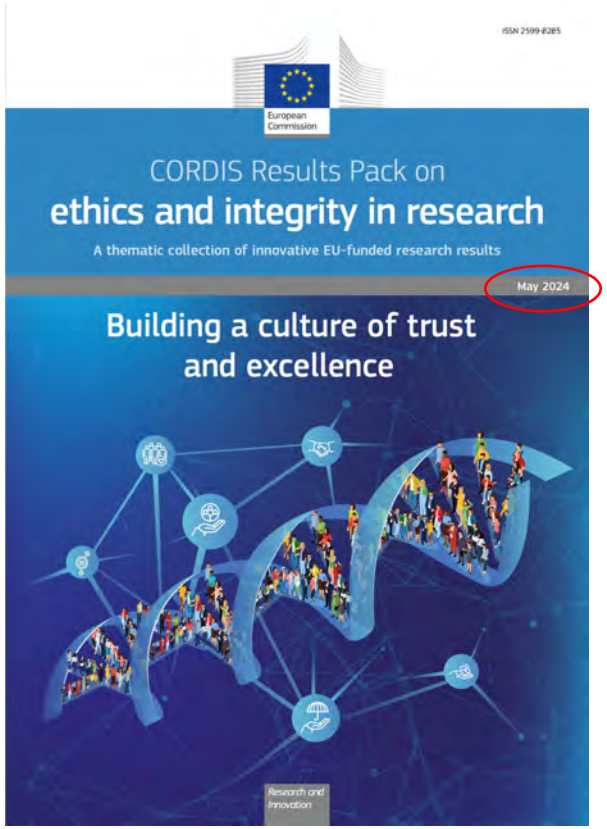
TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

TEAeM - TechEthos Anticipatory ethics Matrix

The combination of the expert reviews and reflection on the prior ethics frameworks have led to the TechEthos Anticipatory ethics Model - TEAeM

Note * denotes a step detailed in ATE+ (Umbrello et al., 2023)

Post TechEthos Impact



European Commission, Directorate-General for Research and Innovation, Publications Office of the European Union, *CORDIS results pack on ethics and integrity in research*, Publications Office of the European Union, 2024, https://data.europa.eu/doi/10.2830/455544

TechEthos receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101006249.

FRAIM: Responsible AI in organisational policy

- ➔ How do **current AI policies** represent responsible AI?
- ➔ Who should be **shaping responsible AI** (in) practice, and how?
- ➔ What do we **mean by 'responsible AI'** anyway?

See more on the [FRAIM website]

Framing Responsible AI Implementation & Management

Arts and Humanities Research Council

BRAID Bridging Responsible AI Divides

British Library

Sheffield City Council

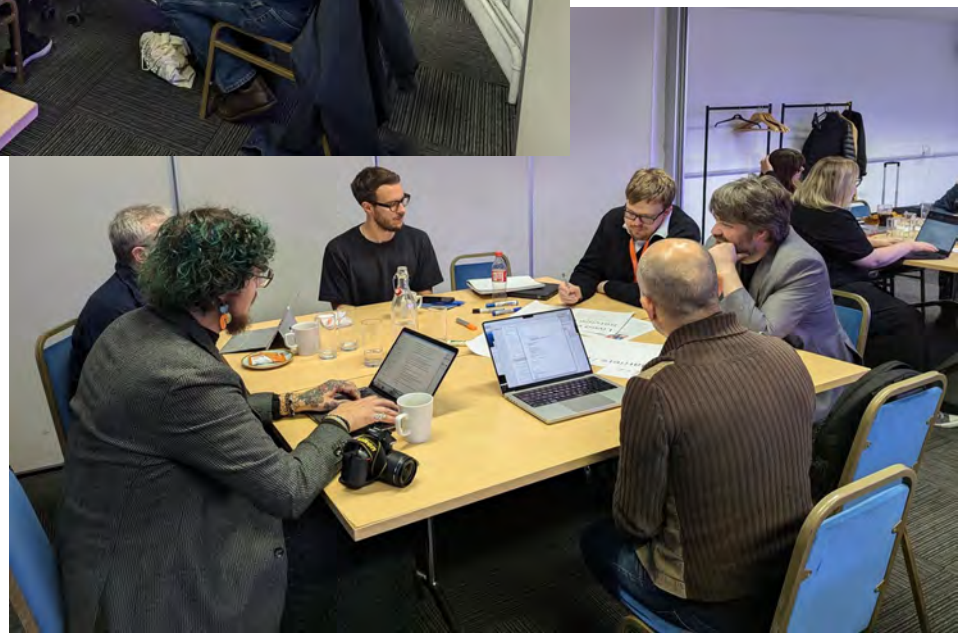
EVIDEN

open data institute

University of Sheffield

The value of co-production in responsible AI research

- Responsibility is **situated** – go where the rubber meets the road
- New insights, research challenges, & **practice-based** ways of thinking
- **Pathways to** (& opportunities for) **impact**



University of Sheffield

Summary: Understanding RAI in organisations

What does RAI mean in organisational practice?

Who are the key stakeholders for RAI in organisations?

WP1: Meta review of RAI literature/resources

- ➔ Underlying questions of ethics/responsibility
- ➔ Stakeholders and audience

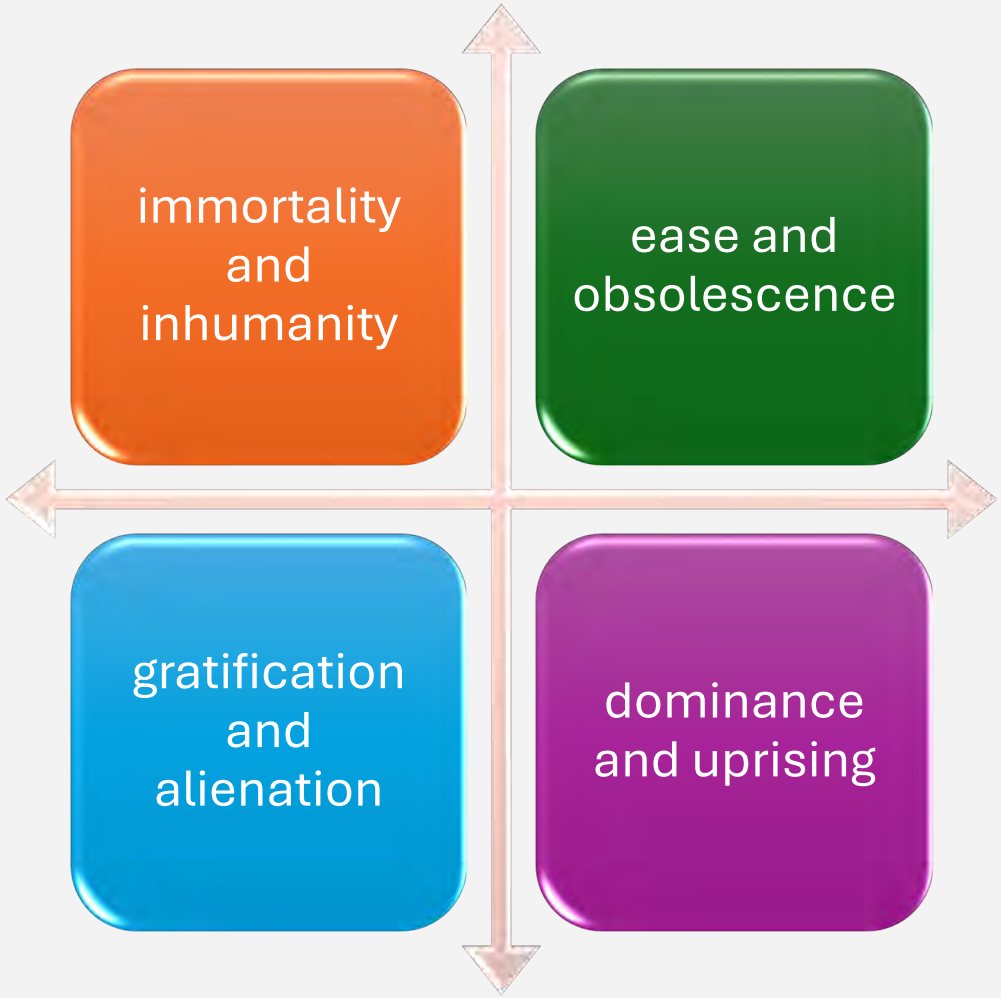
WP2: Interviews within organisations

- ➔ Diversity of RAI roles: implementation, policy, communication, user
- ➔ RAI meaning and practice

WP3: Scoping workshop

- ➔ Needs for evidence base to bring RAI into practice
- ➔ Making RAI more accessible/tangible

Hopes and Fears (Cave and Dihal, 2019)



4 key dichotomous pairs of hopes and fears prevalent in 300 fictional and non-fictional representations of AI:

- Hopes reflect the benefits of maintaining human control over AI;
- Fears arise from the risk of losing control over it.

- Thematic analysis of transcripts from 30 most viewed AI TED talks
- All 8 hopes and fears found – uprising most common, inhumanity and dominance least
- These 'tropes' can be seen as critical tools for fostering shared cultural understandings between experts and lay audiences
  - Communication is highly important in today's politically charged AI environment

Conclusions I

SHERPA

- The ecosystem approach – the SHERPA framework

TechEthos

- Identify previous frameworks: Anticipatory Technology Ethics (ATE) & ATE+, Ethical Impact Assessment (EIA) and Future Studies
- The TechEthos approach integrates the theoretical ethical frameworks with policy documents and empirical data
  - Mapping the policy landscape (rapid review); digital ethnographies; expert interviews and consultation
- Development of the TechEthos Anticipatory ethics Matrix (TEAeM)
  - Enable future and emerging technologies to be able to be developed in a more ethically informed way (i.e. ethics-by-design); to support the ethical governance of the broadest range of future technologies

Conclusions II

So what?
What does this mean for society today, and the future?

- Need to a) be aware of the **potential pitfalls** and other issues, b) make sure that these are not included from the start (where possible), or can be corrected once identified, c) cannot avoid the issues with the "it's too difficult" plea…
- Need to support a more responsible and ethical society by reflecting this in the technologies we develop and use.
  - Who is responsible, can it be just one person, or do we need to adopt some concept of '**distributed responsibility**'?
- Asking the 'right' **questions** of the people and organisations involved in the development and use…
  - Making the argument for the legitimacy of asking these questions…get more people to take it seriously?
- Employ the various **tools** available to create more responsible technologies…before it is too late!
  - Danger of tools being too flexible (possible ethics washing) or too strict (unresponsive to context).
- No single simple solution…
- Focus on the **choices** we have and those we make!
- Pandora's box is open…what is left at the bottom…hope!

Symbiosis vs Terminator...

- Thank You & Questions?