Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

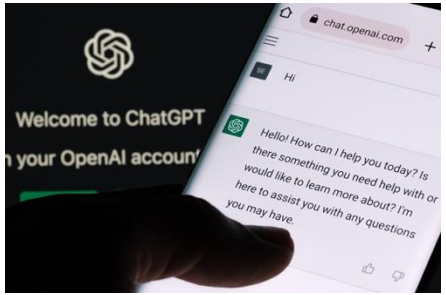
Ethics on the Internet - 1
Technology Ethics, Data and AI

What is the Internet Doing to Me

Delaram Golpayegani
sgolpays@tcd.ie

Thanks to Prof. Dave Lewis

AI is (Nearly) Everywhere



When did you last engage with AI?

How do you typically use AI?

Do you Trust AI?

In what tasks?

Use of AI is not free of risks!

Misinformation

AI-generated fake content

Bias

Example: Gender Bias in Google Translate

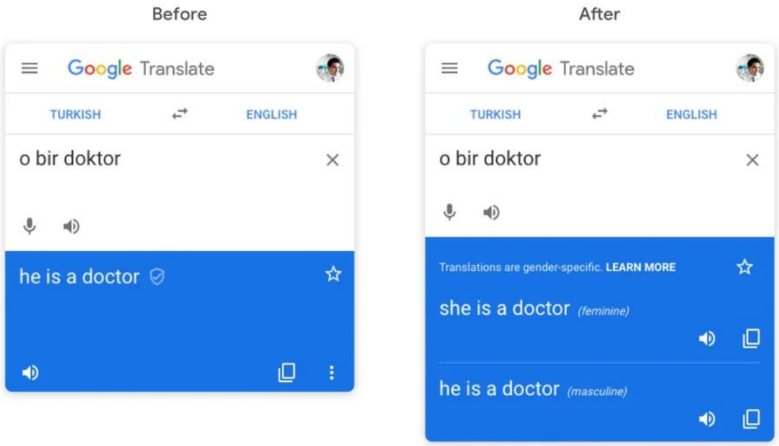
- **Some languages, like Turkish, don't have gender specific pronouns**
- **Google translate has to guess the gender when translating in English**
- **Statements allocating gender to role reveal gender bias**
- **What is the source of this?**
- **Is it a problem?**

Sample Google Translate output:

he is a soldier
she's a teacher
he is a doctor
she is a nurse

https://qz.com/1141122/google-translates-gender-bias-pairs-he-with-hardworking-and-she-with-lazy-and-other-examples/

Mitigation Measure to Reduce Gender Bias in Google Translate



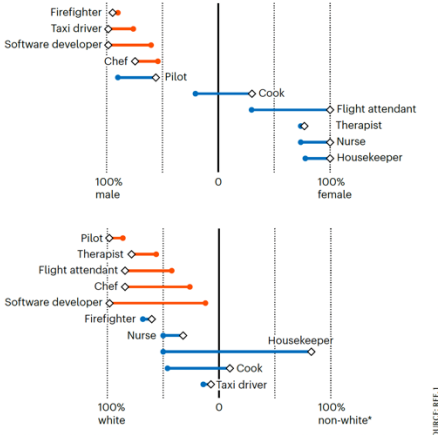
Gender-specific translations on the Google Translate website.

https://blog.google/products/translate/reducing-gender-bias-google-translate/

Gender and Racial Bias in AI-Generated Images

Amplified stereotypes in

AI model outputs



https://www.nature.com/articles/d41586-024-00674-9

https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Power of Big Data – AI Impact on Democracy

Example: Cambridge Analytica

- **Academic research into Psychographics (U. Cambridge) revealed the link between psychological profiles and Facebook profiles**
- **Correlated major psychological types to elements in the social graph: Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism**
- **Cambridge Analytica applied psychographics to help target political ads in 2016 US elections….**

https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump

Algorithmic Power on Behaviour & Worldview

- **“Race to the Bottom … of the Brain Stem” Tristian Harris**
- **70% of YouTube views are based on algorithmic recommendations**
- **Business model maximises video views to maximise ad views**
- **Outrage/fear/anger the most reliable reactions that drive us to keep watching**
- **-> Recommender algorithm inevitably drive us to content that builds outrage to keep us watching**

- Evidence to US Congress: https://www.youtube.com/watch?v=WQMuxNiYoz4
- Agenda: https://humanetech.com/wp-content/uploads/2019/06/Technology-is-Downgrading-Humanity-Let%E2%80%99s-Reverse-That-Trend-Now-1.pdf

Big Data and AI on the Internet

<u>Big Data</u> are extremely large data sets that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions.

- Examples: location traces; social media posts/likes/comments; digital content in the form of text, audio, video; geospatial data; sensor data

<u>Artificial Intelligence (AI)</u> is a family of computational techniques that aim to mimic human capabilities such a learning and problem solving

<u>Machine Learning</u> is an increasing successful form of AI using mathematical models trained on Big Data rather than explicit coding instructions

- Example applications: media recommender systems, speech recognition, face recognition, natural language processing, machine translation, search, predictive data analytics

AI & Data is Mainstreaming Technology Ethics

- **Companies harvest and utilise personal data on a massive scale**
- **Growing concerns about the collection, linking, use and leakage of personal data from mobile devices, bio-sensors, cameras, GPS trackers and social media.**
- **Machine Learning deliver new levels of insights and predictions about an individual's behaviour also feeds increasingly personalised AI-driven interactive digital experiences –Digital Engagement From Ads to Alexa**
- **Individuals and groups struggle to understand the impact of personal information processing**
- **Companies, especially SMEs, often lack the knowledge and expertise needed to address these complex legal and ethical issues.**

2022 - Step Change in AI Capabilities

<u>Language Models</u>

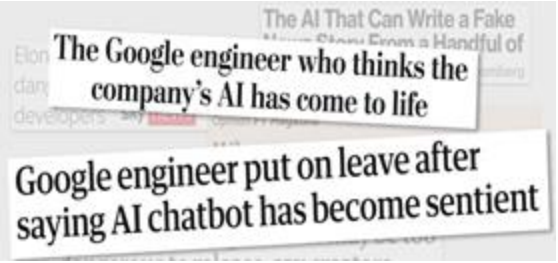
- **Machine learning model that underpin Natural Language Processing tasks**
- **Translation, question-answering, speech recognition, summarization, entity recognition, etc.**

<u>Large Language Models (LLM)</u>

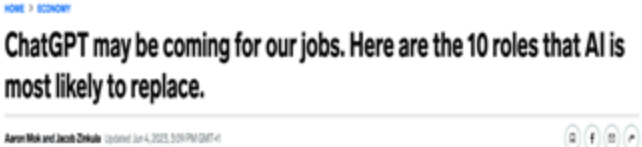
- **Trained on vast content data sets crawled from the Web**
- **Surprised that LLM excel at a <u>wide range</u> of tasks**

<u>Foundational LLMs</u>

- **Models that can be easily adapted to new tasks**
- **Prompt Engineering, Reinforcement Learning from Human Feedback, Model Fine Tuning**



How ChatGPT Managed to Grow Faster Than TikTok or Instagram



Risks: Algorithmic selection of digital content on the Internet

- **Manipulation of individuals or groups,**
- **Diminishing variety that creates biased views and distortion of reality,**
- **Constraints on communication and freedom of expression,**
- **Threats to privacy and data protection rights,**
- **Social discrimination,**
- **Violation of intellectual property rights,**
- **Impact on the human brain and cognitive capacity and**
- **Algorithmic power over human behavior and development.**

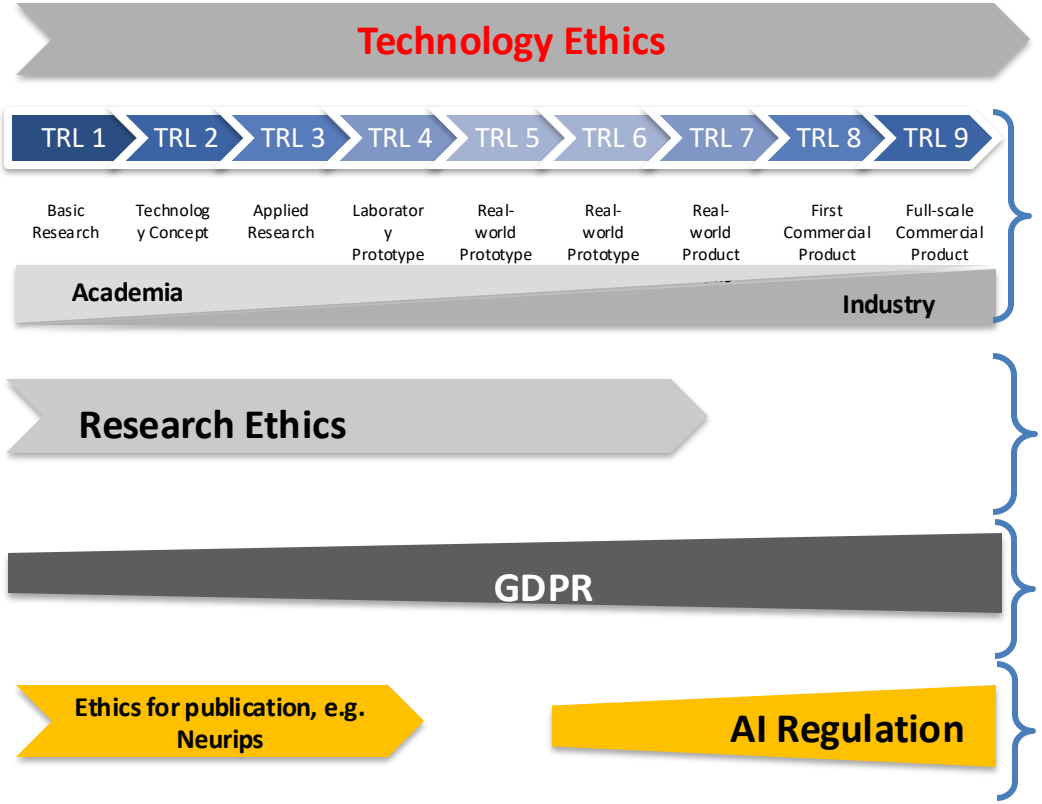
Latzer, M., Hollnbuchner, K., Just, N., & Saurwein, F. (2016). The economics of algorithmic selection on the Internet. Handbook on the Economics of the Internet, (October 2014), pp 395–425. Retrieved from https://doi.org/10.4337/9780857939852.00028

Why Should Digital Tech Innovators be Concerned with Ethics?

- New digital technologies have a profound impact on the way we live, on the relationships we have, on the societal & political processes we engage in.
- **For tech innovators?**
  1. It is good for the image of your business (instrumental goal)
  2. It actually improves the service you provide! (substantive goal)
  3. It is the *good* thing to do, it contributes to your idea of a better society and being a good person (normative goal)
  4. **Law** requires it.

TRL=Technology Readiness Level

Technology Ethics in Context



Technology Ethics concerns how technology impacts society – risks can emerge (and be mitigated) from basic research to market deployment

Research Ethics specifically addresses how research can be conducted without harming human participants

Data protection applies at all stages of technology R&D if using personal data

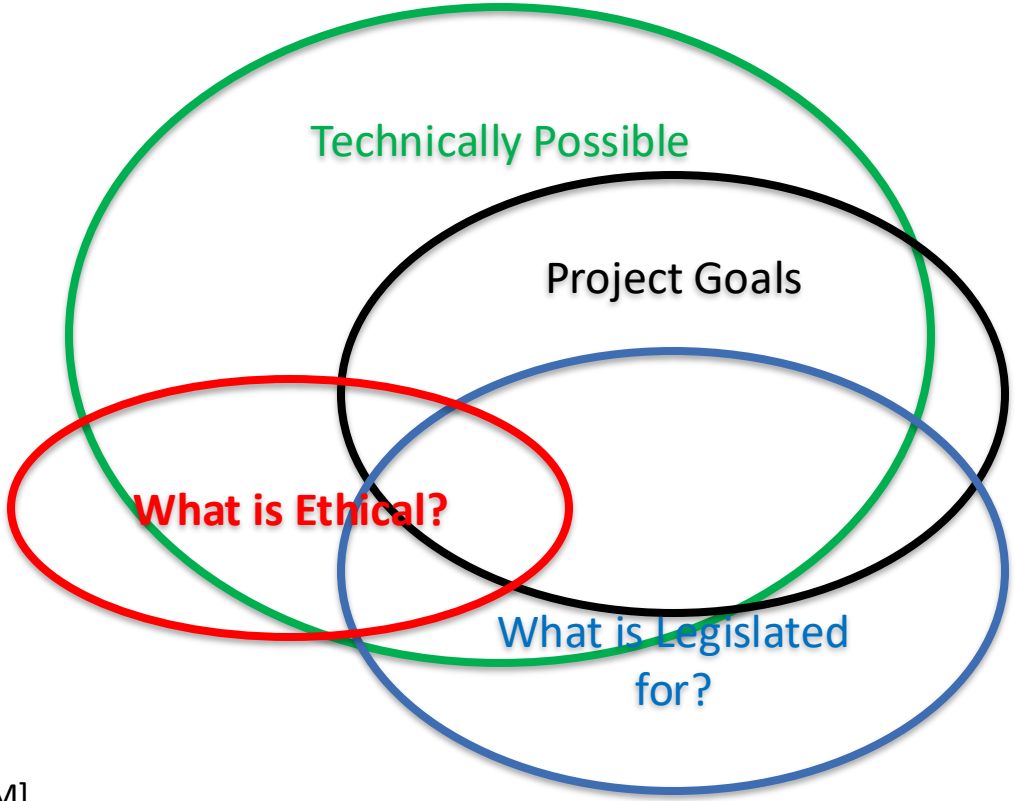
AI and Data bring a new initiatives on ICT R&D ethics and regulation

AI Research and Innovation in the EU:
Who am I accountable to? What should I do? What verdicts can I be subject to?

Career Development

Your Role?	Accountable to Whom?	What must you demonstrate?	Possible Sanction?
Planning Publicly-funded Research	Institution Ethics Committee	Clear research goals, minimise risks to human paerticipants, data protection	Research prevented or delayed
Processing Personal Data (GDPR)	Org's Data protection Office, Data Protection Authorities, Courts	Data Protection: risk assessment, by design, security measures, minimisation, storage limits, subject rights	Public determinations, Legal sanctions, large fines, compensation to subjects
Publishing Research	Peer reviewers/editors	Consideration of professional standards and discipline-specific concerns, e.g. ACM, ACL	Feedback to improve, required corrections,
Seeking Funding	Funding body & its reviewers	Data Management Plan, specific measures in high-risk research areas: clinical, human subject, animals, nuclear, biomed etc.	Funding denied/delayed on ethics/GDPR grounds
Deploying Product or Service (EU AI Act)	External- or Self-Certification Authorities, AND Market Surveillance Authorities, the Courts	GDPR compliance, AI Quality Management System, Impact assessments & Risk Management, technical documentation	Denied product certification, products removed from market, large fines, commercial liability

Ethics in a Technology Development Project



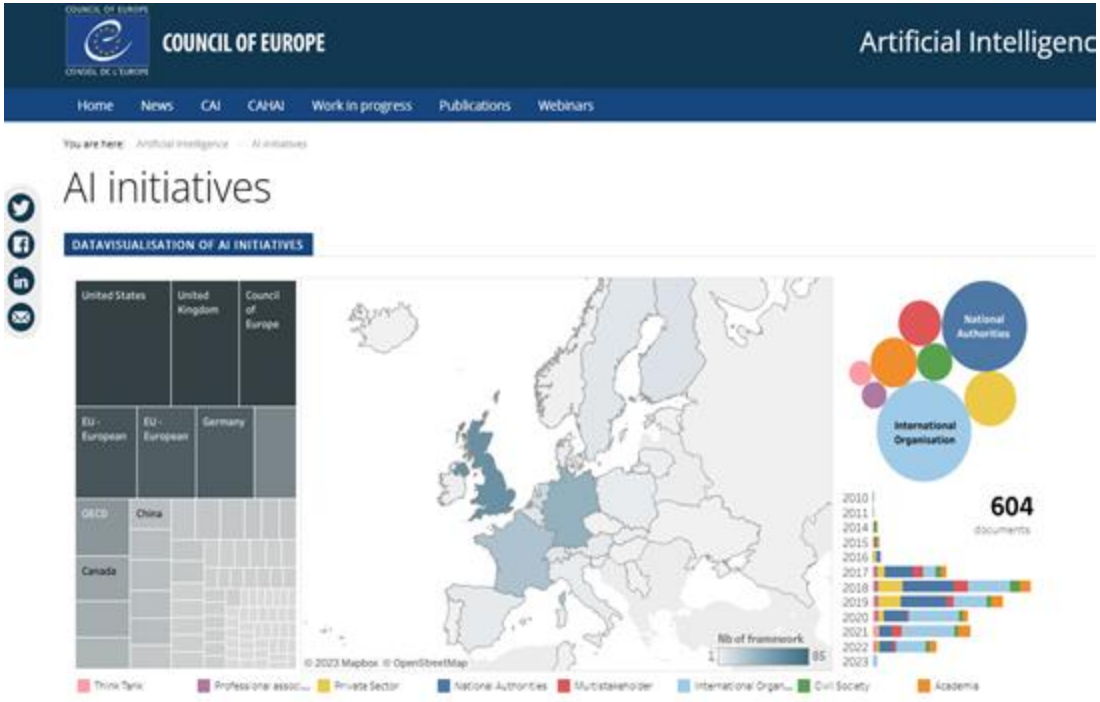
[IBM]

Trustworthy/ Ethical/ Responsible AI

- **Mitigate AI risks and and increase trust and acceptance of the systems**
- **Some characteristics:**
  - Fairness
  - Explainability
  - Accountability
  - Reliability
  - Security

Kaur, Davinder, et al. "Trustworthy artificial intelligence: a review." *ACM computing surveys (CSUR)* 55.2 (2022): 1-38. https://dl.acm.org/doi/full/10.1145/3491209

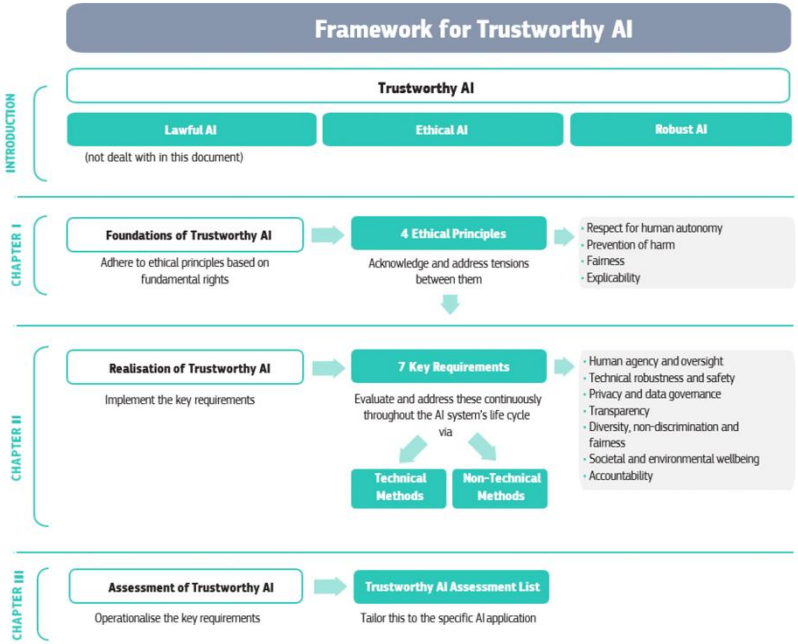
Flourishing of Trustworthy/ Ethical/ Responsible AI initiatives



https://www.coe.int/en/web/artificial-intelligence/national-initiatives

Example of Ethical Principles:

EU Trustworthy AI Guidelines



https://data.europa.eu/doi/10.2759/346720

Example of Ethical Principles:

EU Trustworthy AI Guidelines

https://data.europa.eu/doi/10.2759/346720

EU Ethics Guidelines for Trustworthy AI

Risk Mitigation Methods

- Technical:
  - Architecture,
  - Ethics/privacy-by-design,
  - Explanation,
  - Testing/validation,
  - QoS Indicators
- Non-Technical:
  - Regulation
  - Code of Conduct
  - Standardisation
  - Certification
  - Accountability via Governance Frameworks
  - Education & Awareness
  - Stakeholder Participation
  - Diverse Design Teams

https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

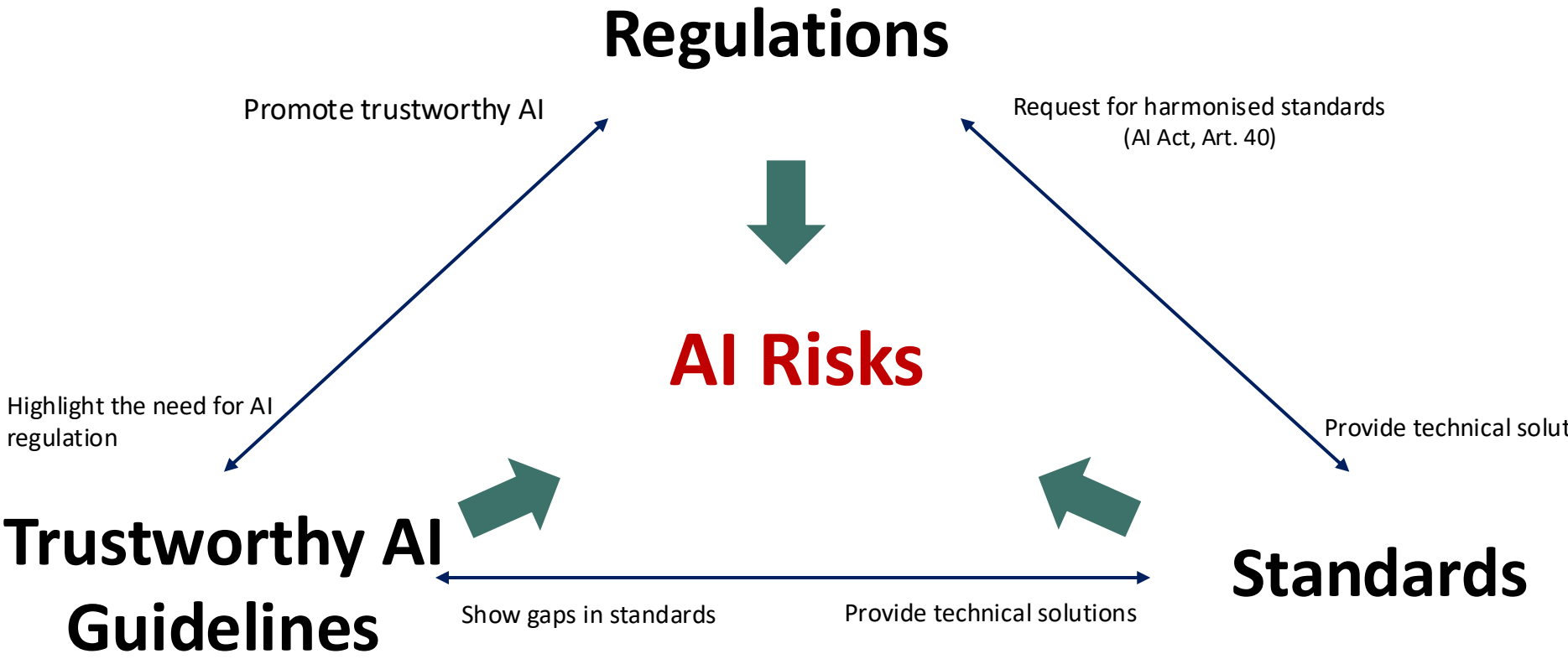
Competing/Converging Sets of Principles

Consensus on principles of

- **Transparency**
- **Justice**
- **Non-maleficence**
- **Responsibility**
- **Privacy**

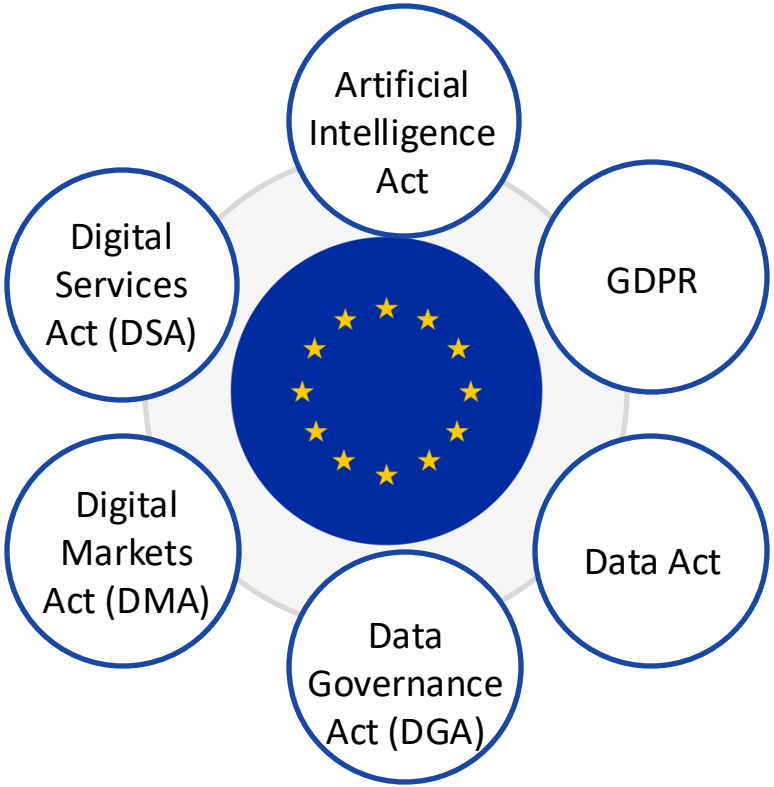
Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2

Dealing with AI Risks



The Big 5+1 EU Digital Regulations

for Data and AI Economy



The EU AI Act

New Rules for

- AI Systems

- GPAI Models **[General Purpose AI]**

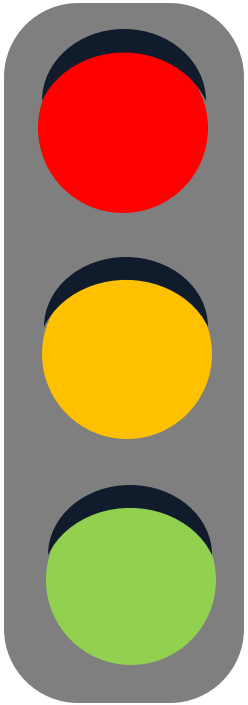
Promotes human-centric & trustworthy AI

Protects against harmful effects of AI on

- **Health**
- **Safety**
- **Fundamental Rights**

The AI Act can be accessed at: http://data.europa.eu/eli/reg/2024/1689/oj

The AI Act Risk-Based Approach



Prohibited

High-Risk

Chapter III, Section 2: Requirements for high-risk AI systems
Chapter III, Section 3: Obligations of providers and deployers of high-risk AI systems and other parties

Non-High-Risk

Art. 95: Codes of Conduct

Art.50: **Transparency** Obligations for providers and deployers of certain AI systems

AI Act's High-Risk AI Systems

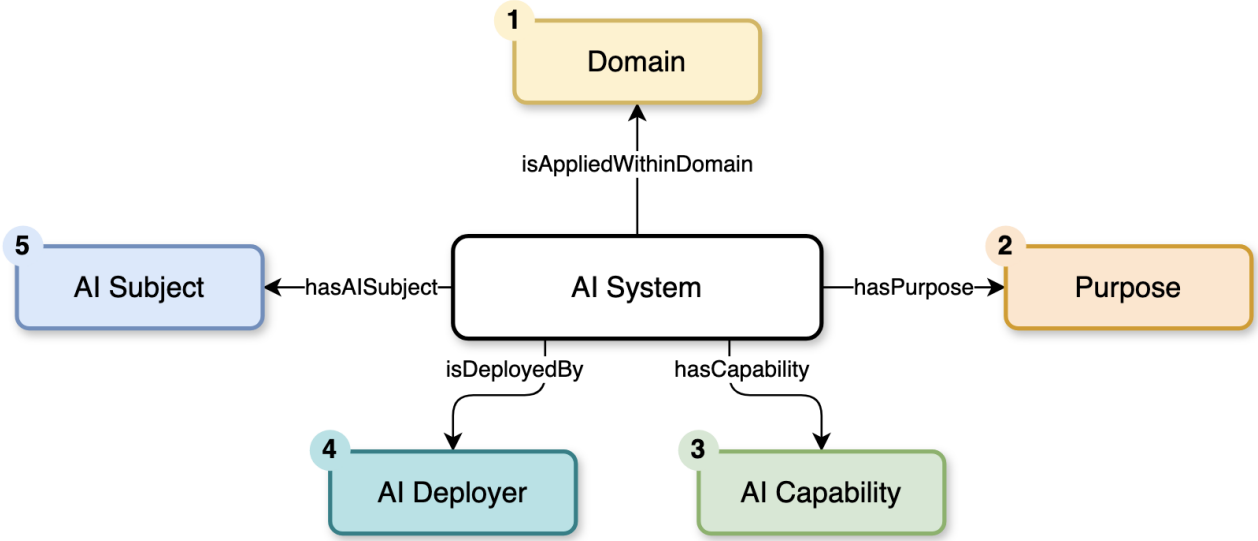
Annex I

- Already regulated areas
- E.g. toys, machinery, medical devices

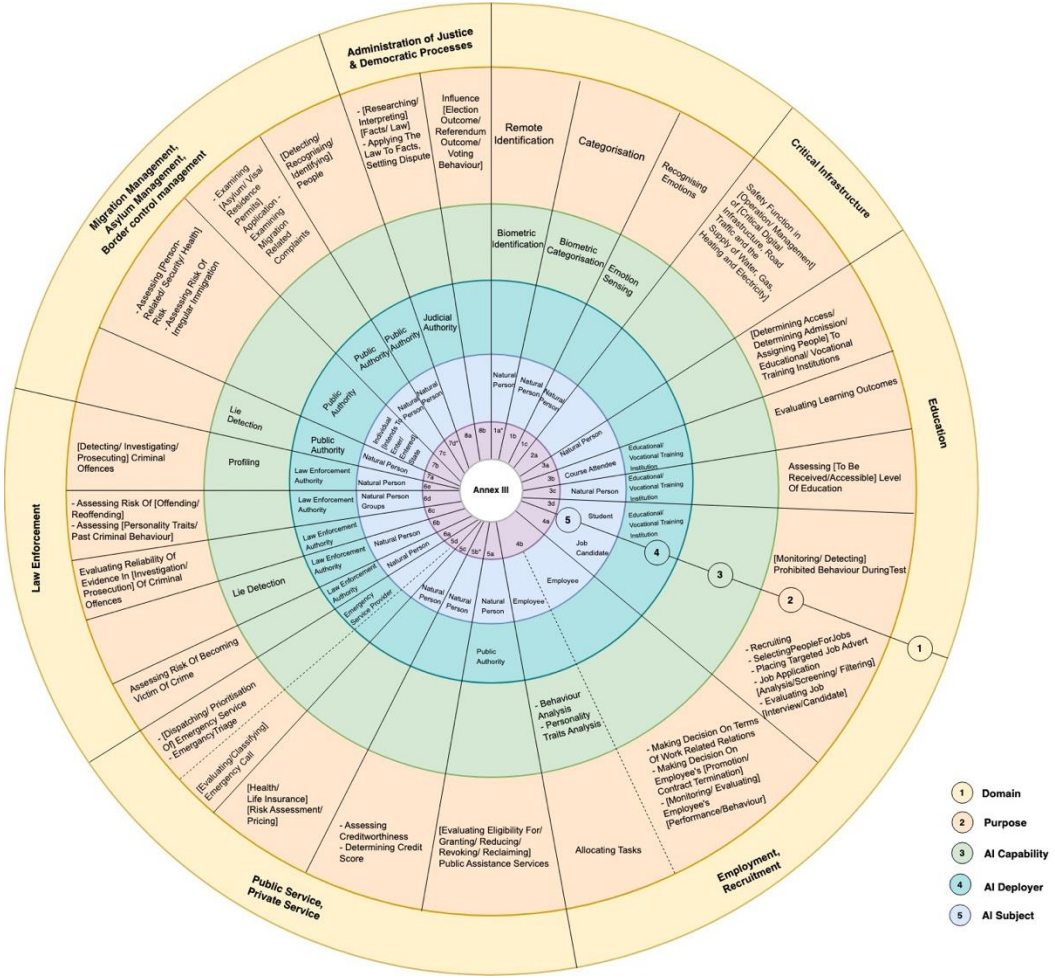
Annex III

- 8 application areas
- E.g. employment, education, law enforcement, migration

5 Core Concepts for Determining Annex III High-Risk AI Systems



Annex III High-Risk Conditions Using the 5 Concepts



Example

(1) In which **Domain** is the AI system used? Law enforcement

(2) What is the **Purpose** of the AI system? Assessing past criminal behaviour

(3) What is the **Capability** of the AI system? Behaviour analysis

(4) Who is the **Deployer** of the AI system? Law enforcement authority

(5) Who is the **AI Subject**? Individuals who are suspected of a crime

The AI system is highly likely to be High-Risk according to Annex III, 6e

Assignment

- Select a high-risk AI application
- Discuss why it is high-risk
- Discuss and document ethical issues using the Ethic Canvas categories (next session)
- Identify risks and mitigations for AI subjects and other affected stakeholders (guided by ISO 26000)
- Summarise how you broke down the work in your group

25% of module mark
Randomly assigned to groups of 5/6
By 8th Nov: Contact group
By 15th Nov: Choose your application
By 28th Nov: Submit report

Where to Look for Inspiration?

AI Incident Repositories

AI, Algorithmic and Automation Incidents and Controversies (AIAAIC) Repository

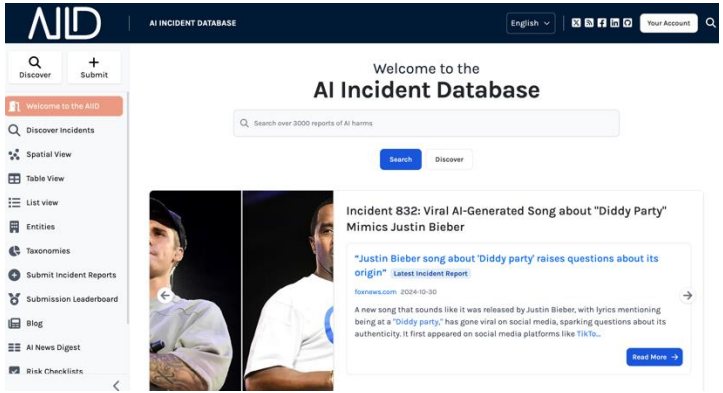
https://www.aiaaic.org/

The AI Incident DataBase (AIID)

https://incidentdatabase.ai/

AIAAIC Repository (beta) [REPORT INCIDENT]

AIAAIC ID#	Headline	Type	Released	Occurred	Country(ies)	Sector(s)	Deployer(s)	Developer(s)	System name(s)	Technolog
AIAAIC1793	Molly Russell, Brianna Ghey chatbots discovered on Cha	Issue		2024	UK	Media/entertainment/t		Character AI	Character AI	Chatbot; Ma
AIAAIC1792	Michael Parkinson AI podcast series sparks ethics contr	Issue		2024	UK	Media/entertainment/t	Deep Fusion Films			Deepfake -
AIAAIC1791	AI search engines promote white supremacism	Issue		2024	Kenya; Pakistan	Politics		Google; Microsoft; Pe	AI Overviews; Copilot;	Chatbot; Ge
AIAAIC1790	UK man jailed for 18 years for creating AI child abuse ima	Incident		2024	UK	Media/entertainment/t	Hugh Nelson	Daz Productions	Daz 3D	Machine lea
AIAAIC1789	Study: OpenAI voice agents can automate phone scams	Issue	2024	2024	USA	Multiple		OpenAI	Realtime API	Bot/intellige
AIAAIC1788	Pensioner loses NZD 224,000 to deepfake scam	Incident		2024	New Zealand	Banking/financial serv				Deepfake -
AIAAIC1787	Synthesia accused of violating AI actors' integrity, trust	Incident		2024	UK	Media/entertainment/t		Synthesia	Synthesia	Machine lea
AIAAIC1786	Whisper speech recognition and transcription	System	2023		USA	Multiple		OpenAI	Whisper	Chatbot; Ge
AIAAIC1785	Study: Whisper AI transcription invents medical treatmen	Issue	2023	2024	USA	Health	Nabla	Nabla; OpenAI	Whisper	Chatbot; Ge
AIAAIC1784	Company uses Marques Brownlee AI voice clone to prom	Incident		2024	USA	Media/entertainment/t	Dot			Deepfake -
AIAAIC1783	Amazon Alexa attributes false facts to fact checking orga	Issue		2024	UK	Media/entertainment/t		Amazon	Amazon Alexa	Virtual assis
AIAAIC1782	Character AI	System	2022		USA	Media/entertainment/t	Character AI	Character AI	Character AI	Chatbot; Ma
AIAAIC1781	Boy commits suicide after relationship with Character AI	Incident		2024	USA	Media/entertainment/t	Sewell Setzer III	Turnitin	Character AI	Chatbot; Ma
AIAAIC1780	AI detectors falsely accuse students of cheating	Incident		2024	USA	Education	Moira Olmsted		Turnitin	Machine lea
AIAAIC1779	Polish radio station replaces humans with AI	Incident		2024	Poland	Media/entertainment/t		OFF Radio Krakow		Machine lea
AIAAIC1778	Researchers find TikTok fails to ban political advertising	Issue		2024	USA	Politics		TikTok	TikTok advertising mar	Machine lea
AIAAIC1777	Experts question opaque Cybercheck AI crime fighting to	Issue		2024	USA	Govt - justice; Govt - p		Global Intelligence	Cybercheck	Machine lea
AIAAIC1776	Dow Jones sues Perplexity AI for copyright abuse	Incident		2024	USA	Media/entertainment/t		Perplexity AI	Perplexity AI	Chatbot; Ma
AIAAIC1775	Tesla sued for using Blade Runner 2049 imagery to launc	Incident		2024	USA	Automotive; Media/en	Elon Musk; Tesla; Warn			Image-to-im
AIAAIC1774	AI nudification bots swamp Telegram	Incident		2024	Croatia; Kosovo	Media/entertainment/t		Alaiksandr Babichau,	ClothOff	Deepfake -
AIAAIC1773	Character AI used to create "disturbing" Jennifer Ann Cle	Incident		2024	USA	Media/entertainment/t		Character AI	Character AI	Chatbot; Ma
AIAAIC1772	NYT orders Perplexity to stop misusing its content	Incident		2024	USA	Media/entertainment/t		Perplexity AI	Perplexity	Chatbot; Ma
AIAAIC1771	AI account recovery scam calls target Gmail users	Issue		2024	Global	Technology				Deepfake -
AIAAIC1770	Common Crawl dataset	Data	2008		USA; Global	Multiple	Common Crawl Founda	Common Crawl Foun	Common Crawl	Database/d

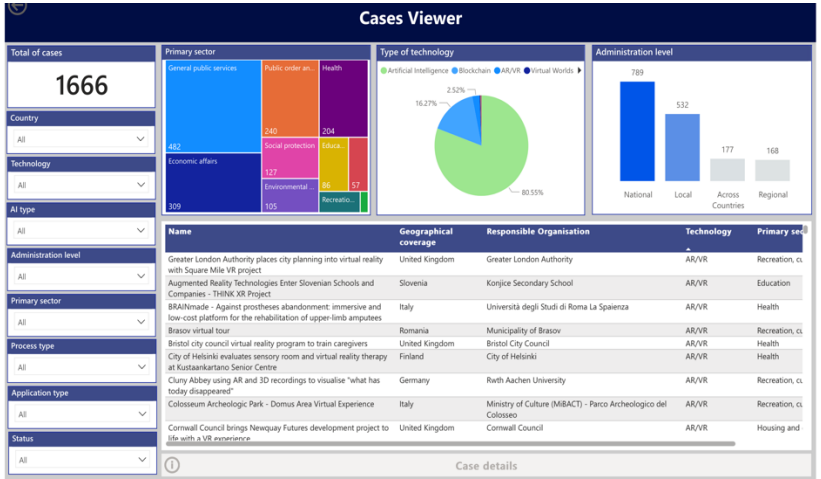


AI in the Public Sector

Public Sector Tech Watch

– Use cases of emerging technologies in the public sector in the EU

https://interoperable-europe.ec.europa.eu/collection/public-sector-tech-watch/cases



High-Risk AI Use Cases

Applied AI's risk database

https://www.appliedai-institute.de/en/risk-classification-database

appliedAI institute for europe

appliedAI Institute | Resources | Educational offers | aai Studio | Events | About us | Career

Risk Class | Transparency Obligation | Enterprise Function | Search database...

Clear All | high-risk | low-risk | prohibited

TITLE OF AI USE CASE / ENTERPRISE FUNCTION	RISK CLASS	TRANSPARENCY OBLIGATION	DETAILS
Medical prognosis Research and development	high-risk	Not verified	
Medical decision support Research and development	high-risk	Not verified	
Predictive Maintenance Example 12 Production and manufacturing	high-risk	Not verified	
Medical prognostic model development Research and development	high-risk	Not verified	

A Tool to Determine whether the System is High-Risk

Is My AI System High-Risk?

A tool to assist you determine whether an AI system is High-Risk according to Annex III of the EU AI Act.

Please fill out the high-risk AI checklist

My AI system Example_HighRisk_6e

is intended to be used in the domain of

Law Enforcement ,

for the purpose of Detecting Criminal Offences ,

has the capability of Profiling .

The system is intended to be deployed by Law Enforcement Authority ,

& the entity who is subjected to its use is Natural Person

Check whether your AI system is high-risk

Your AI system is likely to be High-Risk as per AI Act Annex III-6e: Law enforcement, AI systems intended to be used by or on behalf of law enforcement authorities or by Union institutions, bodies, offices or agencies in support of law enforcement authorities for the profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 in the course of the detection, investigation or prosecution of criminal offences.
Domain:LawEnforcement
Purpose: DetectingCriminalOffences
Capability: Profiling
Deployer: LawEnforcementAuthority
Subject: NaturalPerson

https://regtech.adaptcentre.ie/highrisk

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Thank You