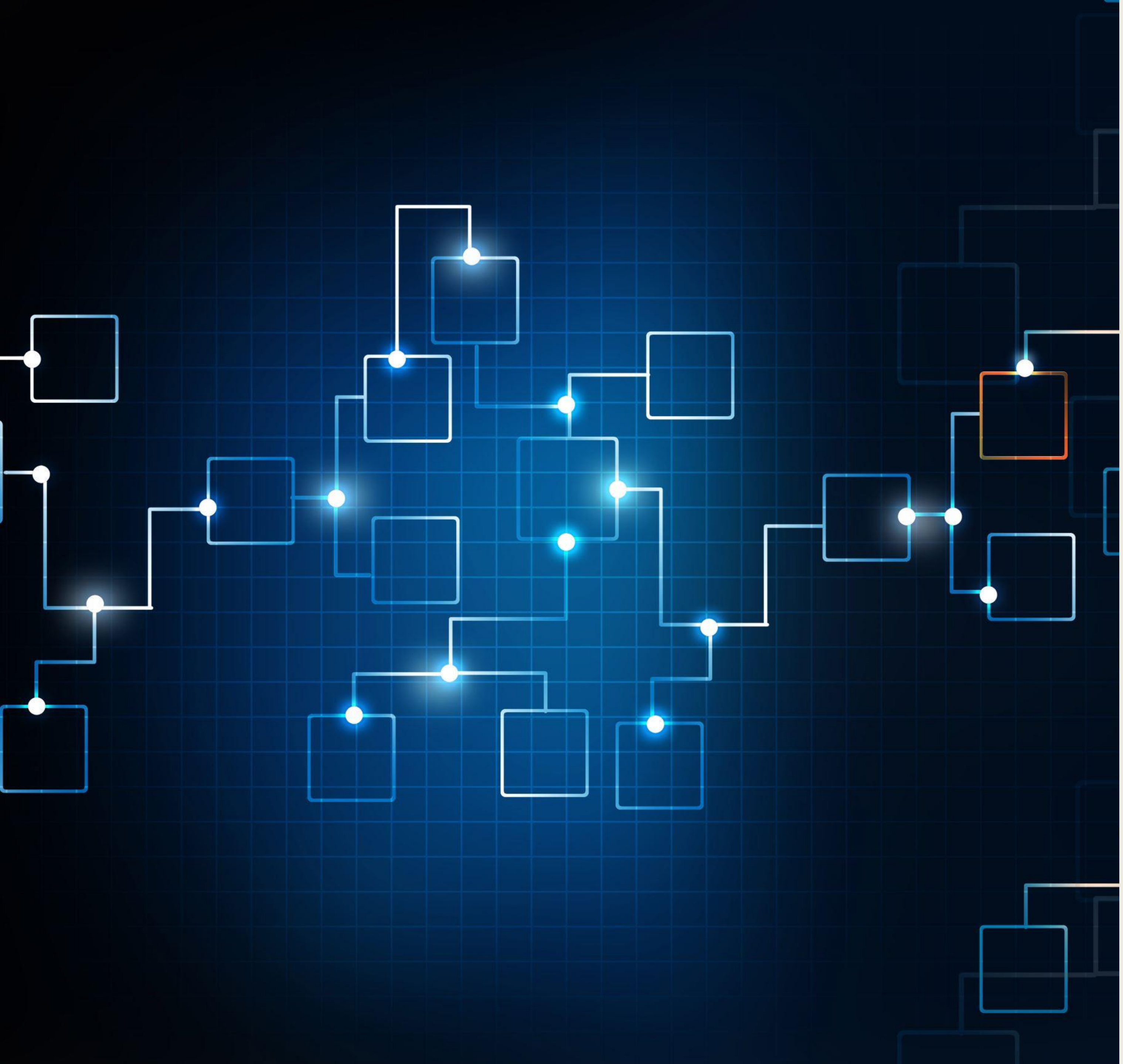


HIT391

MACHINE LEARNING: ADVANCEMENTS AND APPLICATIONS

- **Lecturer:**
- *Dr. Yan Zhang (Danana)*
- *Dr. Al-Amoodi, Abdullah (Sydney)*
- **Email:**
- *yan.zhang@cdu.edu.au*
- *abdullah.al-amoodi@cdu.edu.au*

CHARLES DARWIN UNIVERSITY AUSTRALIA

The Contents are heavily inspired by Sebastian Raschka's course on ML

Week 2: Applications and Ethics Issues of ML

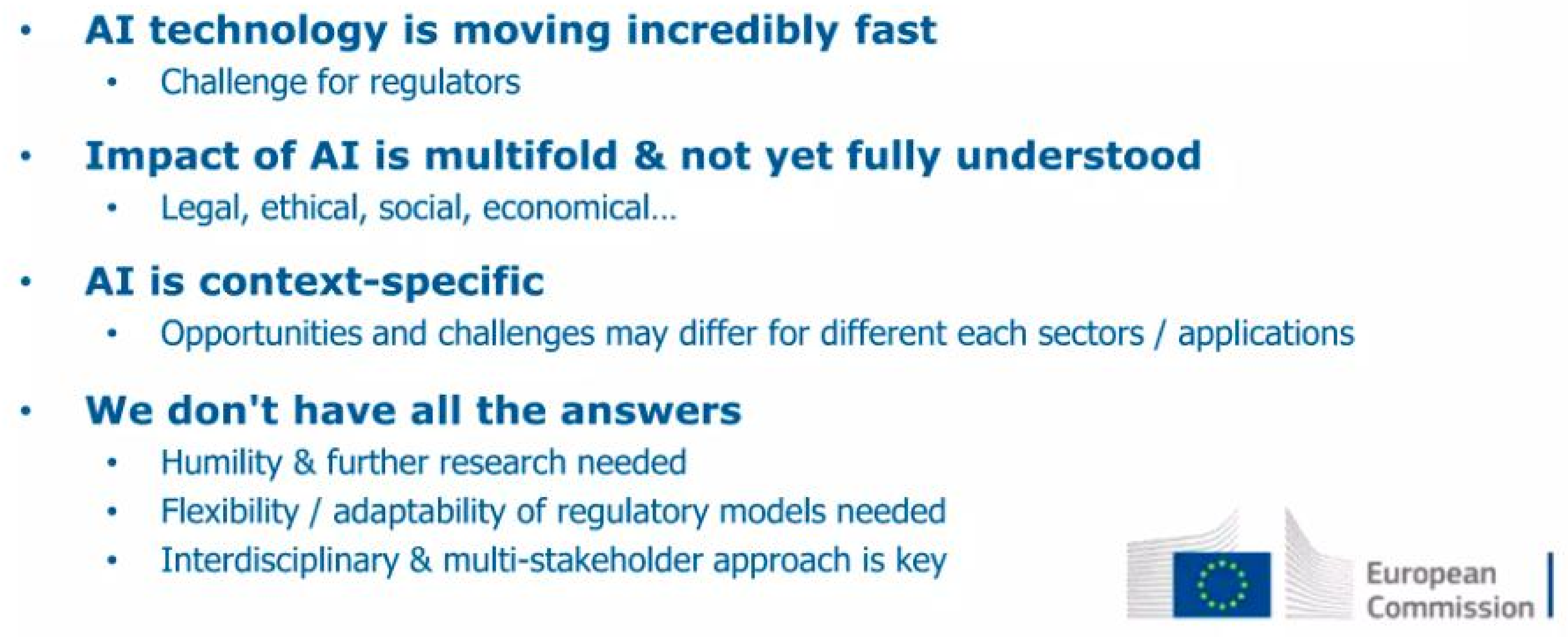
- Learning Outcomes

- Guidelines for Trustworthy AI

Outline

- ❑ Background
- ❑ Ethical Issues of ML
- ❑ Current Legislation
- ❑ Bias of ML

Background



- **AI technology is moving incredibly fast**
  - Challenge for regulators
- **Impact of AI is multifold & not yet fully understood**
  - Legal, ethical, social, economical…
- **AI is context-specific**
  - Opportunities and challenges may differ for different each sectors / applications
- **We don't have all the answers**
  - Humility & further research needed
  - Flexibility / adaptability of regulatory models needed
  - Interdisciplinary & multi-stakeholder approach is key

European Commission

Ethical issues of ML

Ethical Concepts

- What is an ethical issue?

  - Moral issues are those actions which have the **potential to help** or **harm** others or ourselves[1].

- What is an ethical dilemma?

  - A situation in which **a difficult choice** has to be made between two courses of action, **either of which** entails **violating a moral principle**.

Ethics of Technology

- Definition

  - is an interdisciplinary research area concerned with **all moral and ethical** aspects of technology in society. (Luppicini, 2008)

- It views society and technology as interrelated and aims to:

  - use technology ethically.

  - prevent misuses.

  - guide new technological advances.

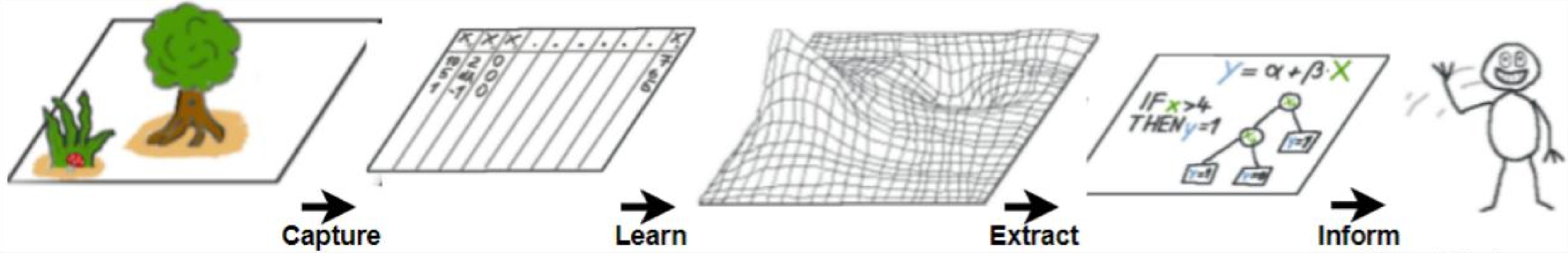
  - benefit society.

Current Ethical Issues

- Fairness
- Discrimination
- Ownership
- Transparency/Interpretability
- Privacy
- Responsibility
- Anonymity
- Confidentiality
- Identity
- Reputation
- …

What is Interpretability?

World Data Black-box Model Interpretability Methods Humans



@towardsdatascience.com

Interpretability = Transparency + Explainability

Model Specific

Model Agnostic

Local

Global

Ethical ML Issues

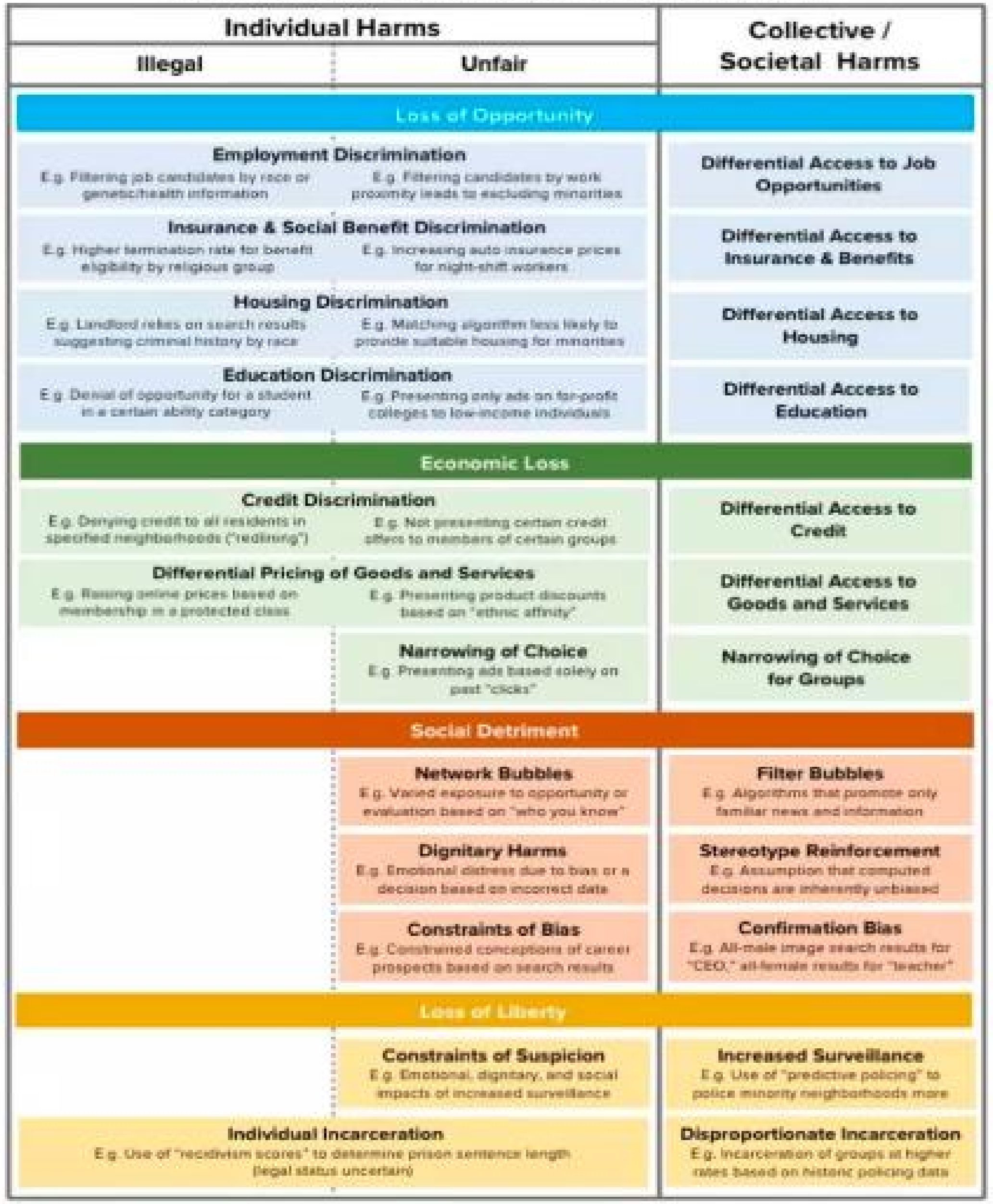
- Autonomous cars
- Autonomous weapons

  - meaningful human control?
- Internet of things (IoT)
- Personalized medicine (genetic information)
- Social Credit System (China)

  - just another credit score?
- Each technical innovation brings risks and benefits.

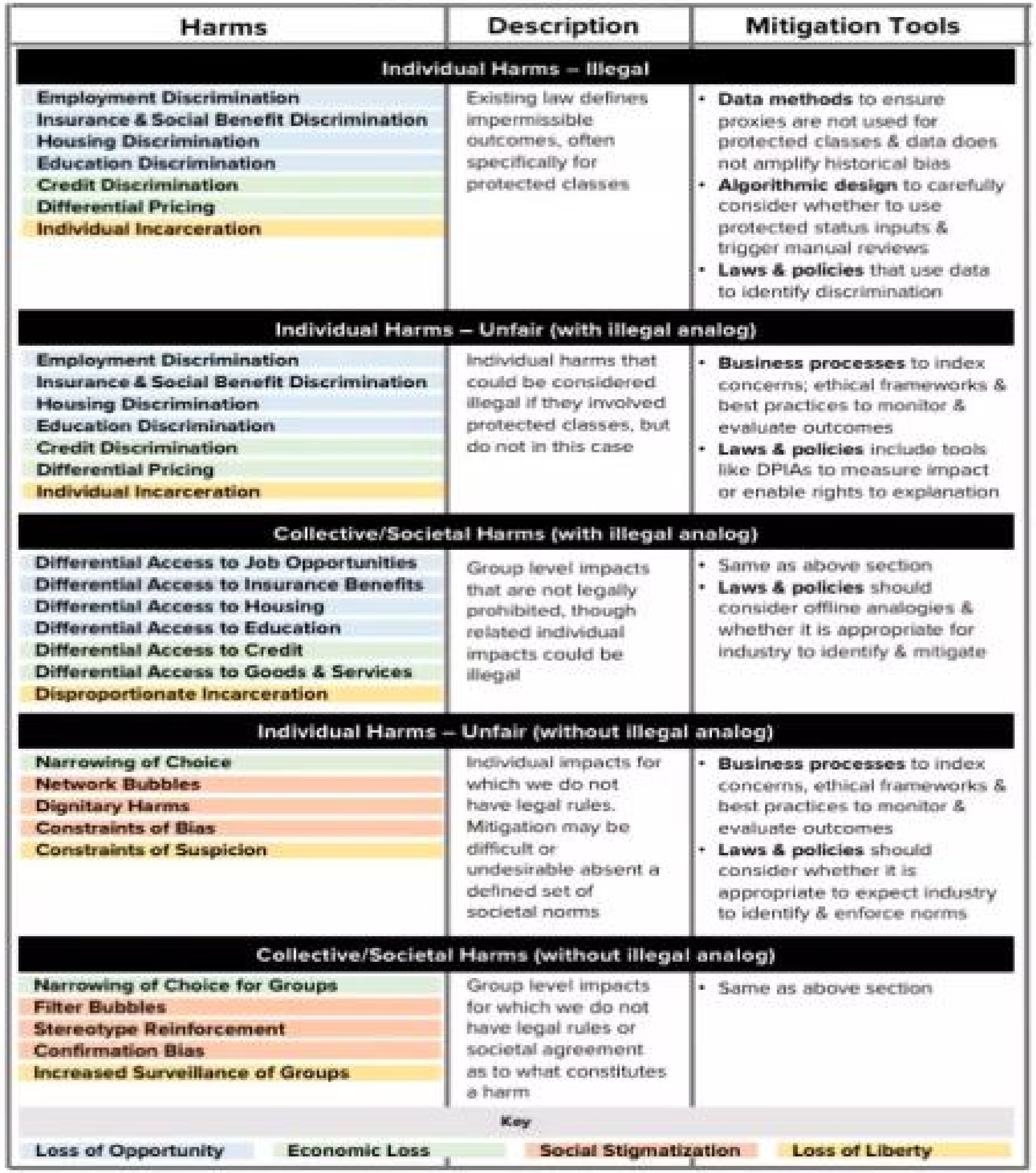
What are some challenges of Automated Decision Making?

Potential Harms from Automated Decision-Making



Individual Harms		Collective / Societal Harms
Illegal	Unfair	
Loss of Opportunity		
Employment Discrimination E.g. Filtering job candidates by race or genetic/health information	E.g. Filtering candidates by work proximity leads to excluding minorities	Differential Access to Job Opportunities
Insurance & Social Benefit Discrimination E.g. Higher termination rate for benefit eligibility by religious group	E.g. Increasing auto insurance prices for night-shift workers	Differential Access to Insurance & Benefits
Housing Discrimination E.g. Landlord relies on search results suggesting criminal history by race	E.g. Matching algorithm less likely to provide suitable housing for minorities	Differential Access to Housing
Education Discrimination E.g. Denial of opportunity for a student in a certain ability category	E.g. Presenting only ads on for-profit colleges to low-income individuals	Differential Access to Education
Economic Loss		
Credit Discrimination E.g. Denying credit to all residents in specified neighborhoods ("redlining")	E.g. Not presenting certain credit offers to members of certain groups	Differential Access to Credit
Differential Pricing of Goods and Services E.g. Raising online prices based on membership in a protected class	E.g. Presenting product discounts based on "ethnic affinity"	Differential Access to Goods and Services
	Narrowing of Choice E.g. Presenting ads based solely on past "clicks"	Narrowing of Choice for Groups
Social Detriment		
	Network Bubbles E.g. Varied exposure to opportunity or evaluation based on "who you know"	**Filter Bubbles** E.g. Algorithms that promote only familiar news and information
	Dignitary Harms E.g. Emotional distress due to bias or a decision based on incorrect data	**Stereotype Reinforcement** E.g. Assumption that computed decisions are inherently unbiased
	Constraints of Bias E.g. Constrained conceptions of career prospects based on search results	**Confirmation Bias** E.g. All-male image search results for "CEO," all-female results for "teacher"
Loss of Liberty		
	Constraints of Suspicion E.g. Emotional, dignitary, and social impacts of increased surveillance	**Increased Surveillance** E.g. Use of "predictive policing" to police minority neighborhoods more
Individual Incarceration E.g. Use of "recidivism scores" to determine prison sentence length (legal status uncertain)		**Disproportionate Incarceration** E.g. Incarceration of groups at higher rates based on historic policing data

Potential Mitigation Sets



Harms	Description	Mitigation Tools
Individual Harms – Illegal		
Employment Discrimination Insurance & Social Benefit Discrimination Housing Discrimination Education Discrimination Credit Discrimination Differential Pricing Individual Incarceration	Existing law defines impermissible outcomes, often specifically for protected classes	• **Data methods** to ensure proxies are not used for protected classes & data does not amplify historical bias • **Algorithmic design** to carefully consider whether to use protected status inputs & trigger manual reviews • **Laws & policies** that use data to identify discrimination
Individual Harms – Unfair (with illegal analog)		
Employment Discrimination Insurance & Social Benefit Discrimination Housing Discrimination Education Discrimination Credit Discrimination Differential Pricing Individual Incarceration	Individual harms that could be considered illegal if they involved protected classes, but do not in this case	• **Business processes** to index concerns; ethical frameworks & best practices to monitor & evaluate outcomes • **Laws & policies** include tools like DPIAs to measure impact or enable rights to explanation
Collective/Societal Harms (with illegal analog)		
Differential Access to Job Opportunities Differential Access to Insurance Benefits Differential Access to Housing Differential Access to Education Differential Access to Credit Differential Access to Goods & Services Disproportionate Incarceration	Group level impacts that are not legally prohibited, though related individual impacts could be illegal	• Same as above section • **Laws & policies** should consider offline analogies & whether it is appropriate for industry to identify & mitigate
Individual Harms – Unfair (without illegal analog)		
Narrowing of Choice Network Bubbles Dignitary Harms Constraints of Bias Constraints of Suspicion	Individual impacts for which we do not have legal rules. Mitigation may be difficult or undesirable absent a defined set of societal norms	• **Business processes** to index concerns, ethical frameworks & best practices to monitor & evaluate outcomes • **Laws & policies** should consider whether it is appropriate to expect industry to identify & enforce norms
Collective/Societal Harms (without illegal analog)		
Narrowing of Choice for Groups Filter Bubbles Stereotype Reinforcement Confirmation Bias Increased Surveillance of Groups	Group level impacts for which we do not have legal rules or societal agreement as to what constitutes a harm	• Same as above section

Key

Loss of Opportunity	Economic Loss	Social Stigmatization	Loss of Liberty

Big Brother Policing – Ethics?

Source: Multiple Internet Search

Every breath you take
Every move you make
Every bond you break
Every step you take
I'll be watching you
Every single day
Every word you say
Every game you play
Every night you stay
I'll be watching you
Oh can't you see
You belong to me
My poor heart aches
With every step you take
Every move you make
Every vow you break
Every smile you fake
Every claim you stake
I'll be watching you
Since you've gone I been lost without a trac
I dream at night I can only see your face
I look around but it's you I can't replace
I feel so cold and I long for your embrace
I keep crying baby, baby, please
Oh can't you see
You belong to me
My poor heart aches
With every step you take
Every move you make
Every vow you break
Every smile you fake
Every claim you stake
I'll be watching you
Every move you make
Every step you take
I'll be watching you
I'll be watching you

- Lyrics from **POLICE** Song

Current Legislation

Legislation

Different legislative approaches:

- Europe
- Australia

Legislation of EU

- Privacy.

  - a fundamental human right; Europe has a long tradition of privacy legislation.

- Strict EU privacy law – applied to all industries.

- **General Data Protection Regulation (GDPR)** legislation, 2016

  - GDPR is applicable as of May 25th, 2018 in all member states to harmonize **data privacy** laws across Europe.

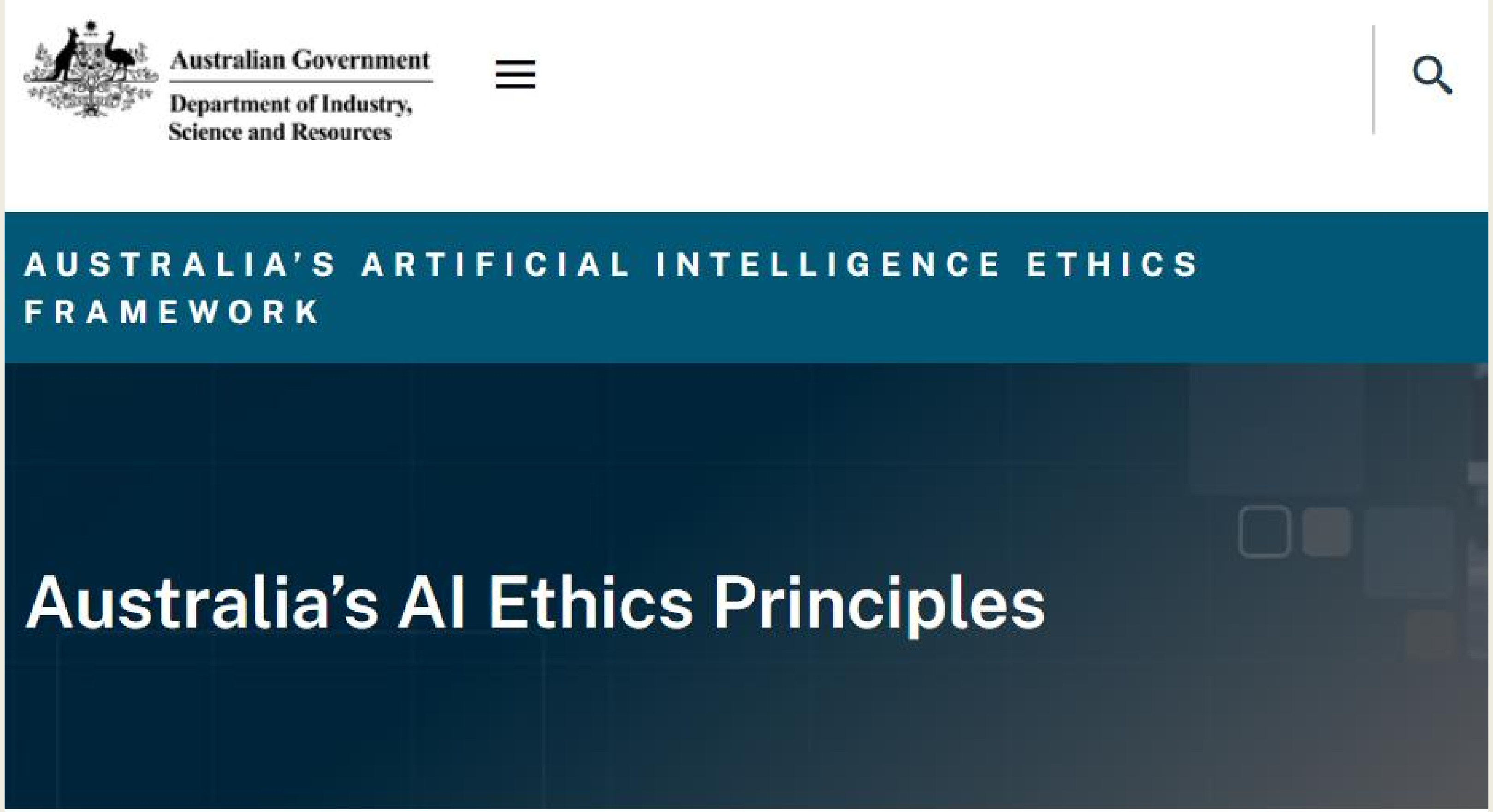
- Less business-friendly environment.

- EU regulations lead to a conflict with US IT corporations. A new special tax on big tech (under discussion 2018-19).

Australia

- Australia's AI Ethics Principles

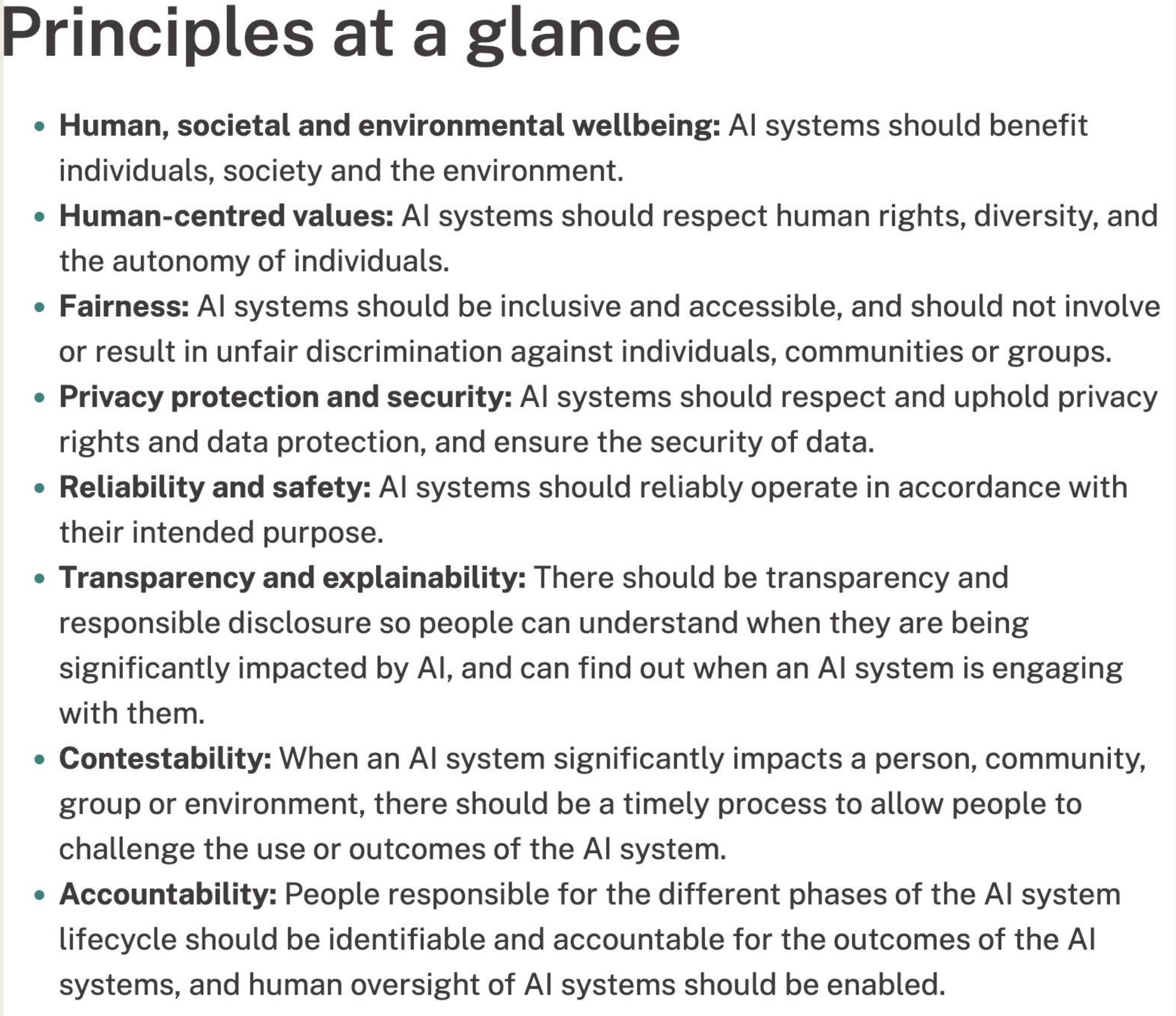
- Australia's 8 Artificial Intelligence (AI) Ethics Principles are designed to ensure AI is safe, secure and reliable.



Australian Government
Department of Industry, Science and Resources

AUSTRALIA'S ARTIFICIAL INTELLIGENCE ETHICS FRAMEWORK

Australia's AI Ethics Principles



Principles at a glance

- **Human, societal and environmental wellbeing:** AI systems should benefit individuals, society and the environment.
- **Human-centred values:** AI systems should respect human rights, diversity, and the autonomy of individuals.
- **Fairness:** AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.
- **Privacy protection and security:** AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.
- **Reliability and safety:** AI systems should reliably operate in accordance with their intended purpose.
- **Transparency and explainability:** There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.
- **Contestability:** When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.
- **Accountability:** People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

European Union

- Document made public on 8 April 2019

European Commission

INDEPENDENT

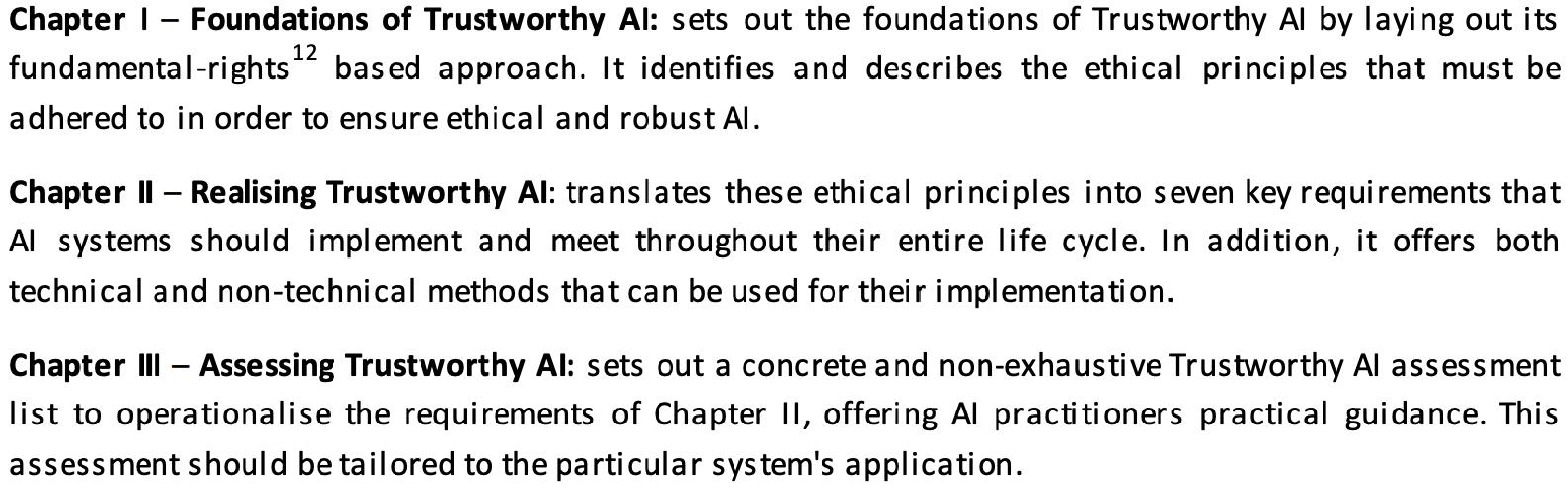
HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

SET UP BY THE EUROPEAN COMMISSION

ETHICS GUIDELINES FOR TRUSTWORTHY AI

European Union

- Ethics guidelines for trustworthy AI

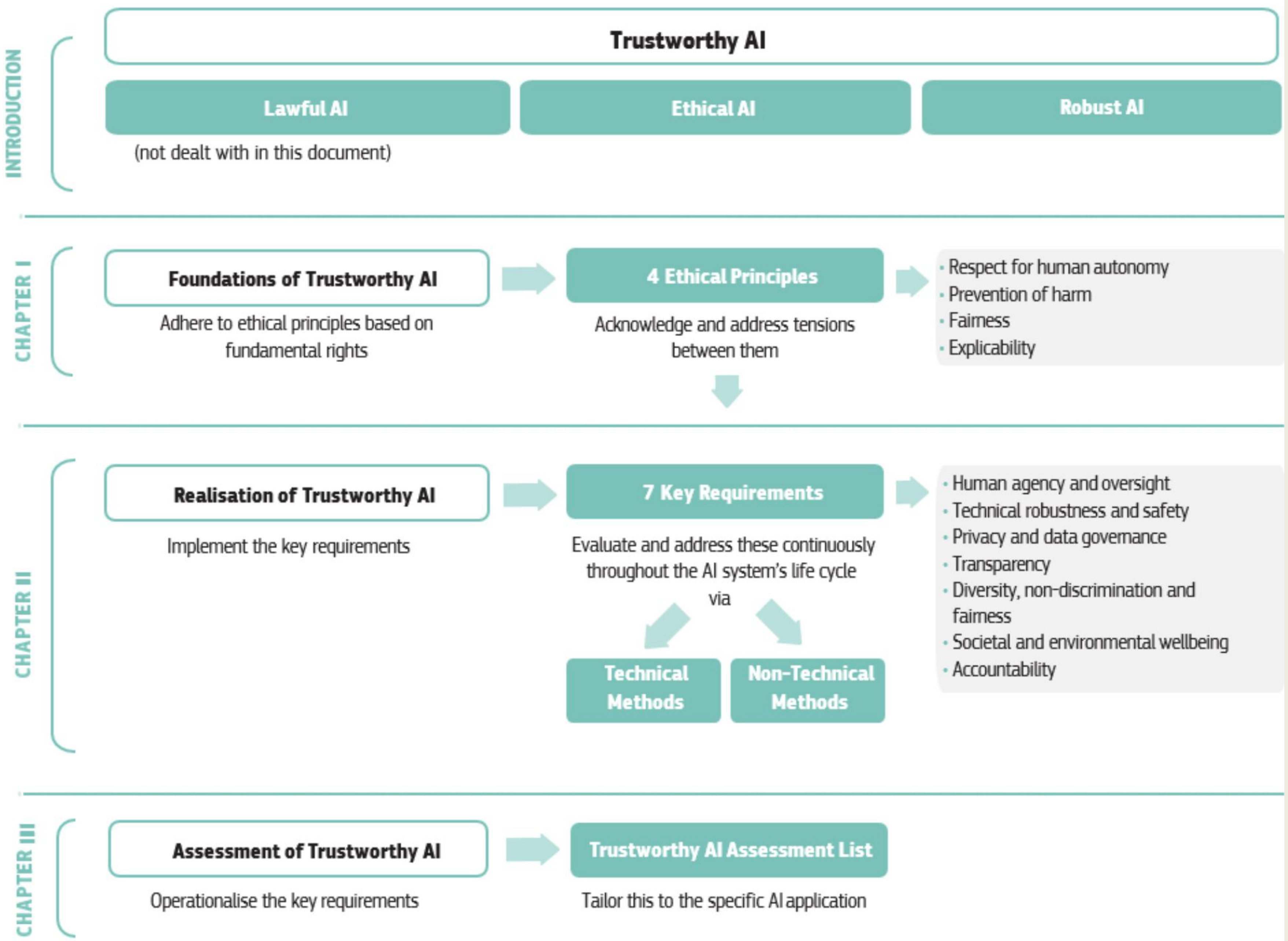


Chapter I – Foundations of Trustworthy AI: sets out the foundations of Trustworthy AI by laying out its fundamental-rights[12] based approach. It identifies and describes the ethical principles that must be adhered to in order to ensure ethical and robust AI.

Chapter II – Realising Trustworthy AI: translates these ethical principles into seven key requirements that AI systems should implement and meet throughout their entire life cycle. In addition, it offers both technical and non-technical methods that can be used for their implementation.

Chapter III – Assessing Trustworthy AI: sets out a concrete and non-exhaustive Trustworthy AI assessment list to operationalise the requirements of Chapter II, offering AI practitioners practical guidance. This assessment should be tailored to the particular system's application.

Framework for Trustworthy AI



INTRODUCTION

Trustworthy AI

Lawful AI

(not dealt with in this document)

Ethical AI

Robust AI

CHAPTER I

Foundations of Trustworthy AI

Adhere to ethical principles based on fundamental rights

4 Ethical Principles

Acknowledge and address tensions between them

- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability

CHAPTER II

Realisation of Trustworthy AI

Implement the key requirements

7 Key Requirements

Evaluate and address these continuously throughout the AI system's life cycle via

Technical Methods

Non-Technical Methods

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental wellbeing
- Accountability

CHAPTER III

Assessment of Trustworthy AI

Operationalise the key requirements

Trustworthy AI Assessment List

Tailor this to the specific AI application

Bias in ML

Current Issues in ML

- Hard to explain the final decision to users since **ML systems** look like **black boxes** (DL algorithms).
- Some of the current ML algorithms behave **unfair**.
- AI/ML systems need to be used by professionals outside engineering/math communities.
- AI/ML systems should be incorporated into **social and legal systems**.

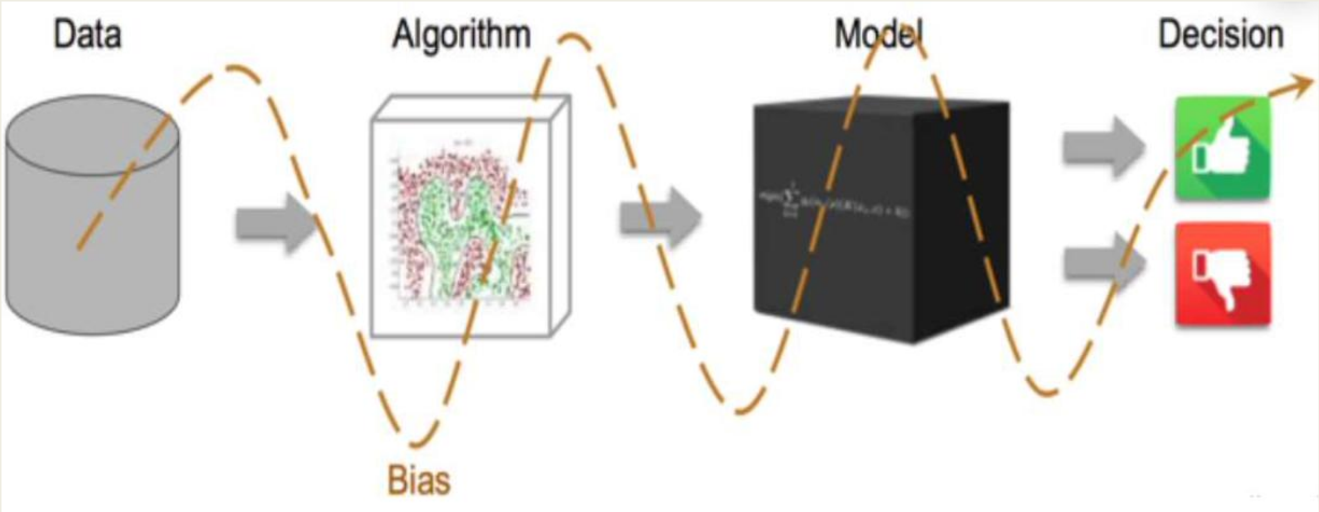
Bias

- Bias (Legal):
  - A **personal** and often **unreasoned judgement** for or against one side in a dispute. (Essential 25000 English Law Dictionary)
- Bias (Statistics):
  - A systematic inaccuracy in data due to the characteristics of the process employed in the **creation, collection, manipulation, and presentation of data**, or due to **faulty sample** design of the estimating technique. (Dauer, F. W. (1989). Critical thinking: An introduction to reasoning)

Bias in ML

Types of biases we are interested in:

- Algorithmic bias, feature or model selection
- Data bias, **biased or irrelevant data**, garbage-in / garbage-out
- Model bias, **Interpretability/Transparency** of DS/ML systems



Problem of Bias in Precision Medicine

Bias in Datasets

Datasets can become unintentionally biased through

- a lack of cohort diversity
- technical processes of data collection and cleaning
- the specific incorporation of electronic health record data.

Bias in Outcomes

The outcomes of precision medicine research can be discriminatory in many ways

- too much focus on individual responsibility for health
- the marginalization of those population groups with lower health literacy or in less resourced areas
- the potential to shift the accepted forms of biomedical research.

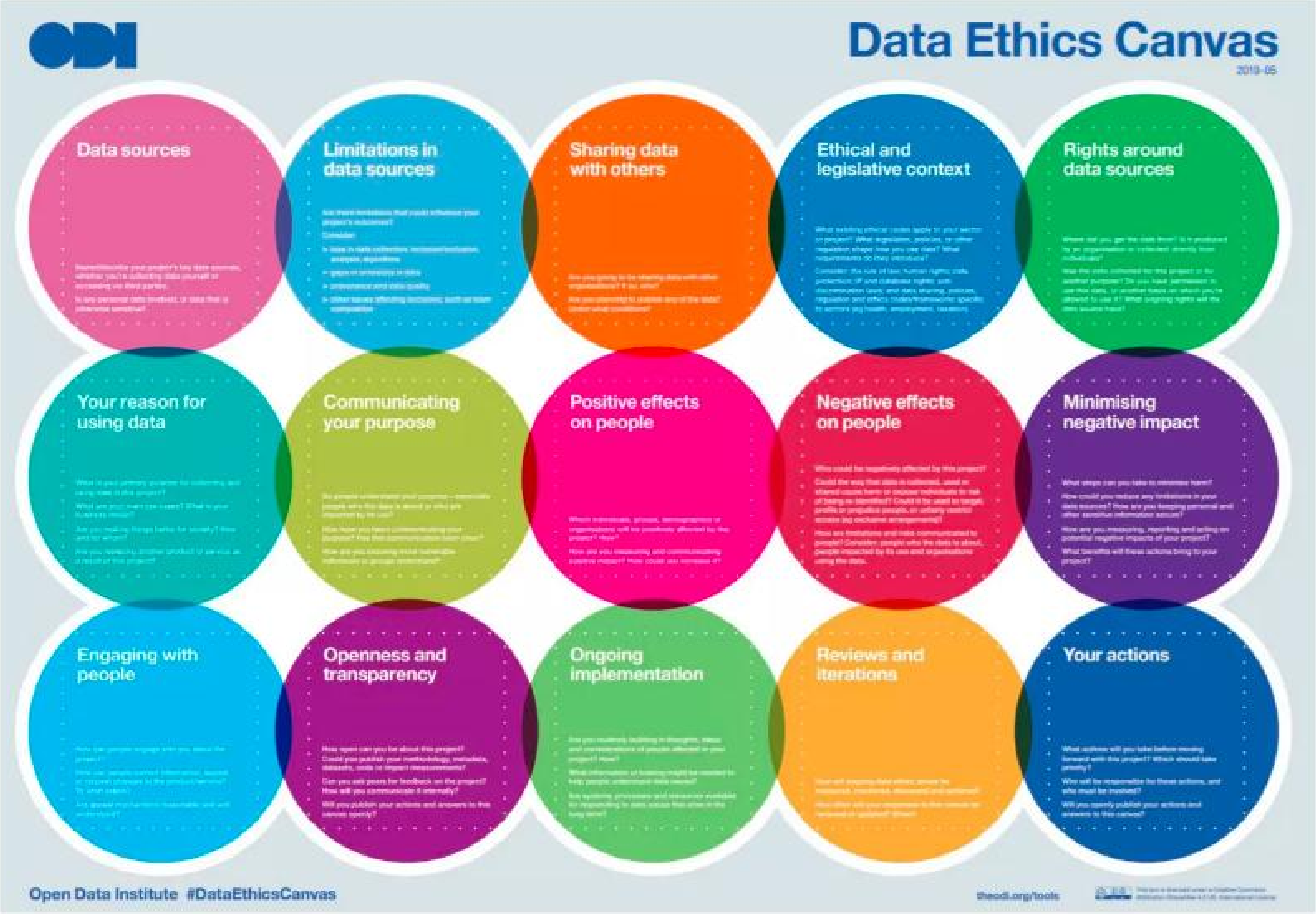
Source : Data&Society – Fairness in Precision Medicine

THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Image Source: Machine Learning, XKCD

How can we handle Data Ethics?

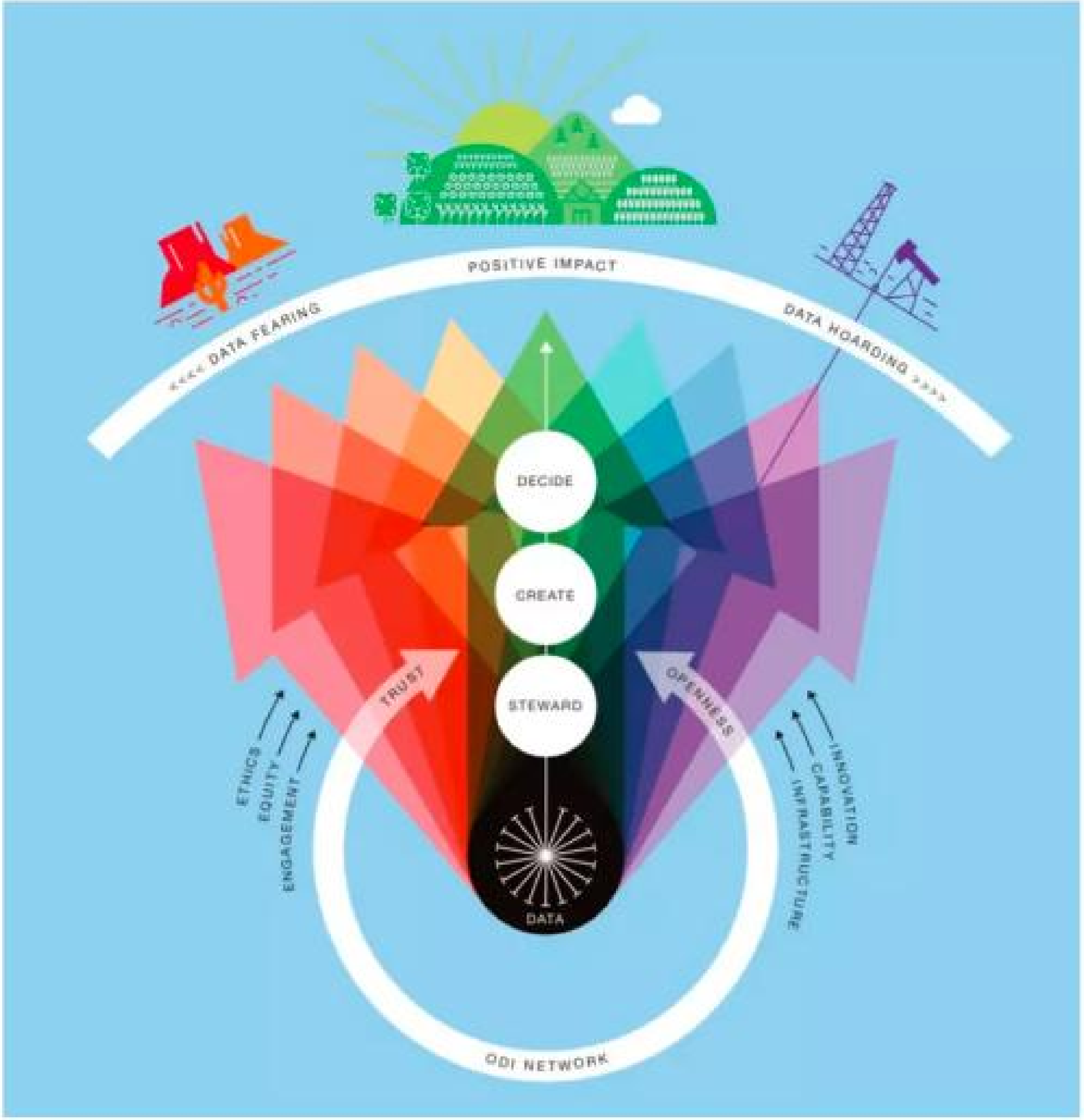
Have a Data Ethics Governance Framework


ODI
Data Ethics Canvas
2019-05
Data sources
Limitations in data sources
Sharing data with others
Ethical and legislative context
Rights around data sources
Your reason for using data
Communicating your purpose
Positive effects on people
Negative effects on people
Minimising negative impact
Engaging with people
Openness and transparency
Ongoing implementation
Reviews and iterations
Your actions
Open Data Institute #DataEthicsCanvas
theodi.org/tools

- Data sources
- Limitations in data sources
- Sharing data with others
- Ethical and legislative context
- Rights around data sources
- Your reason for using data
- Communicating your purpose
- Positive effects on people
- Negative effects on people
- Minimising negative impact
- Engaging with people
- Openness and transparency
- Ongoing implementation
- Reviews and iterations
- Your actions

Source: The ODI

Use the Data Governance to drive innovation ethically



Use the data with an aim to make a positive impact

Decide

- Your approach

Create

- Strategize and Create your Data Governance Strategy

Steward

- Steward and maintain the data Ethically

Use

- Innovation / Capability / Infrastructure
- Ethics / Equity / Engagement
- Leverage Trust / Openness / Network with Standardization bodies such as theODI.org / Fast.AI etc,

Source: The ODI

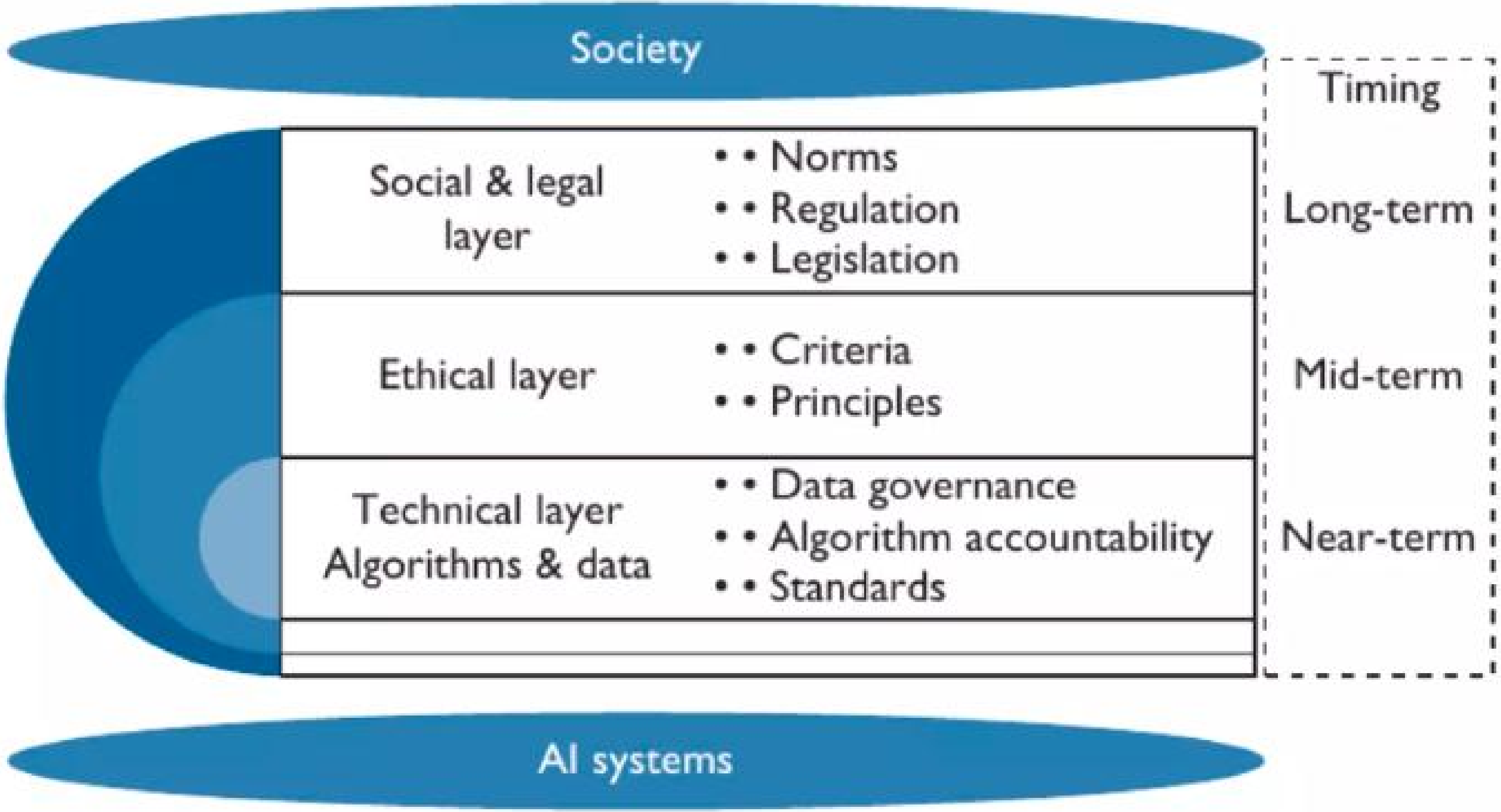
AI Governance Model – A Layered View*

AI-based systems are "black boxes," resulting in massive information asymmetries between the developers of such systems and consumers and policymakers. How do we address this?

AI Blackbox issues

How do we handle issues such as ?

- Justice and equality
- Use of force
- Safety and certification
- Privacy
- Displacement of labor and taxation
- Information asymmetries
- Finding normative consensus
- Government mismatches

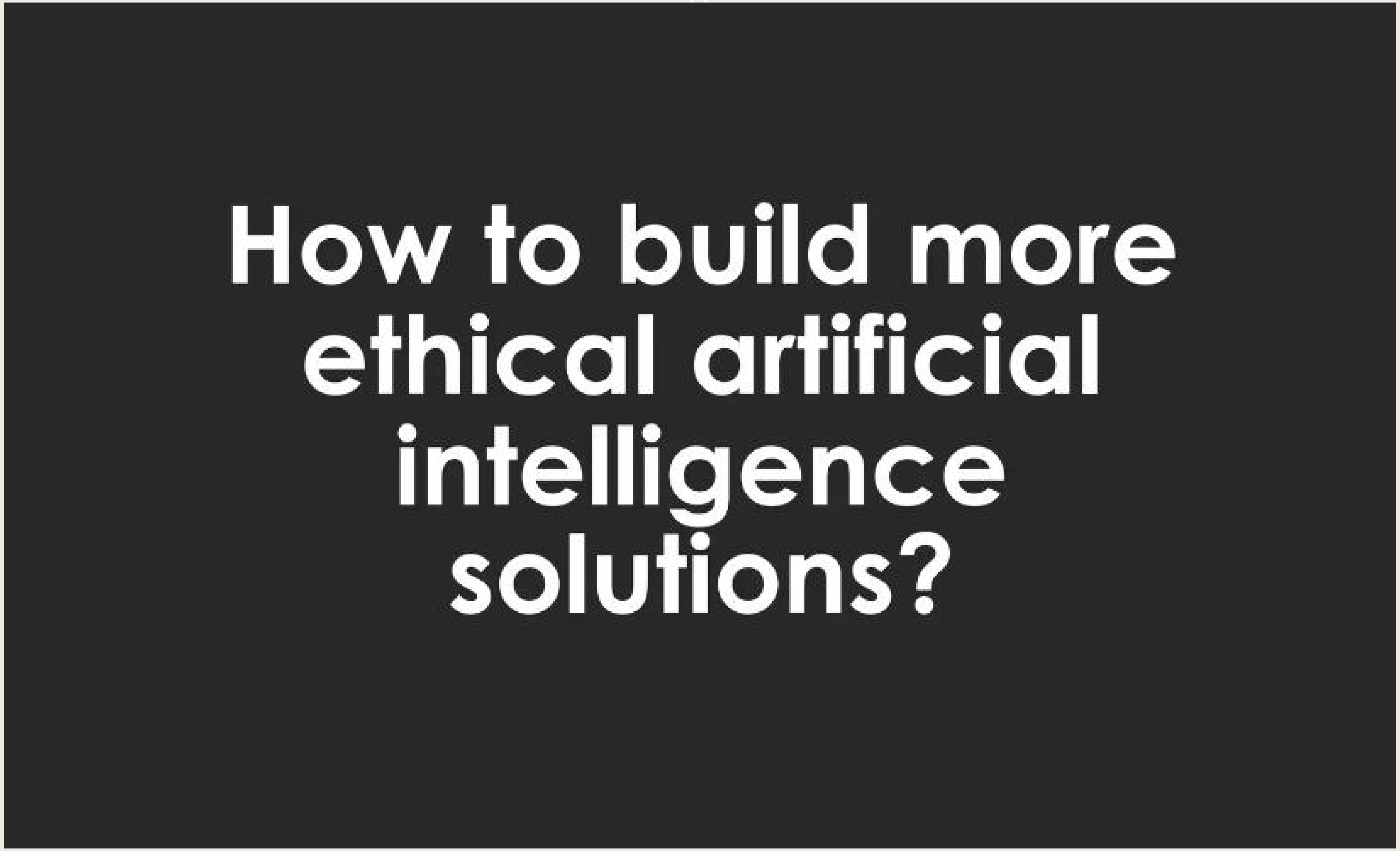


Society

		Timing
Social & legal layer	• • Norms • • Regulation • • Legislation	Long-term
Ethical layer	• • Criteria • • Principles	Mid-term
Technical layer Algorithms & data	• • Data governance • • Algorithm accountability • • Standards	Near-term

AI systems

Source: Urs Gasser and Virgilio A.F. Almeida https://dash.harvard.edu/bitstream/handle/1/34390353/w6gov-18-LATEX.pdf?sequence=1



How to build more ethical artificial intelligence solutions?

@slideshare

Assignment 1 Case Studies

What is Bias in Machine Learning Model and How they can cause Ethical Issues in AI?

What are Black Box Machine Learning Models and How they can cause Ethical Issues in AI?

Ethics of Artificial Intelligence in Medicine

@slideshare

References

- https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework
- https://www.slideshare.net/AndrewDoyle12/ethics-of-artificial-intelligence-in-medicine?from_search=3
- https://www.slideshare.net/krahman/ethics-in-the-use-of-data-ai?from_search=4
- https://www.slideshare.net/vladimirkanchev/ethical-issues-in-machine-learning-algorithms-part-2-140369359