The Alan Turing Institute

AI Ethics and Governance

Day 2: AI Harms and Values

22/11/2022

Professor David Leslie
Director of Ethics and Responsible Innovation

OVERVIEW

Introduction to Practical Ethics

01

AI Harms & Values

02

AI Sustainability

03

Fairness, Bias Mitigation & Accountability

04

Explainability, CARE & Act principles

05

OVERVIEW

01 Introduction to Practical Ethics

02 AI Harms & Values

03 AI Sustainability

04 Fairness, Bias Mitigation & Accountability

05 Explainability, CARE & Act principles

CONTENTS

Day 1 recap

Bioethics and HR

01 AI Harms

Q&A

Lunch break

Activity 1: Thinking about AI harms

02 AI Values

Q&A

Activity 2: Relating to values

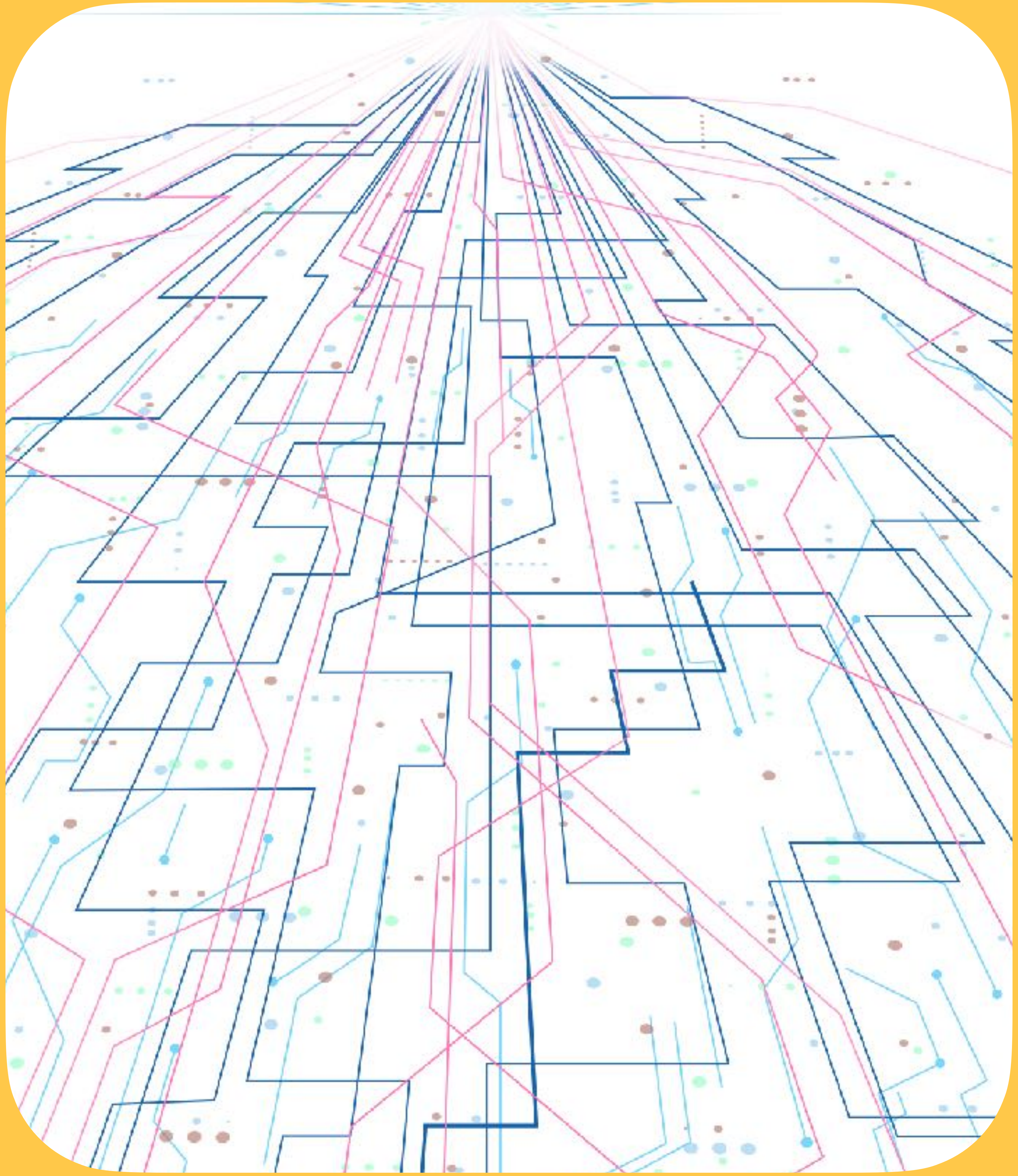
Recap from Day 1

Practical Ethics

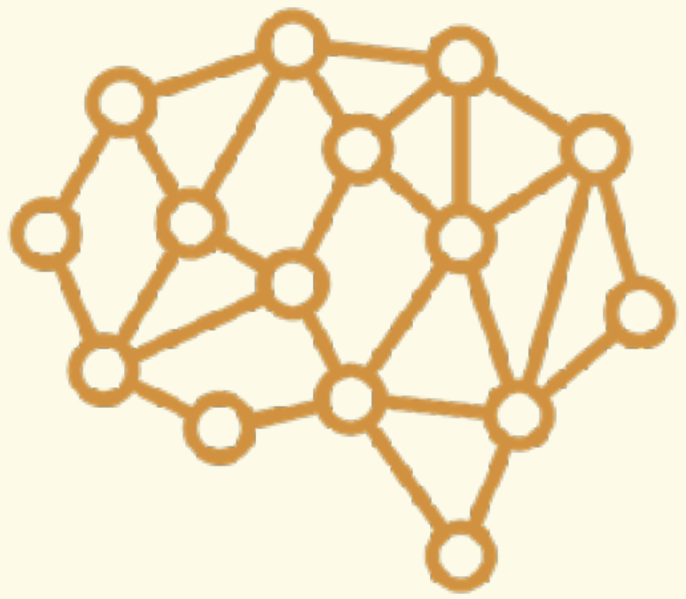
- Why do we care about ethics?
- Approaches to metaethics
- Procedural approach to ethics
- Normative theories:
  - Consequentialism
  - Deontology
  - Virtue ethics
  - Biocentric ethics

01

AI HARMS



What is AI?



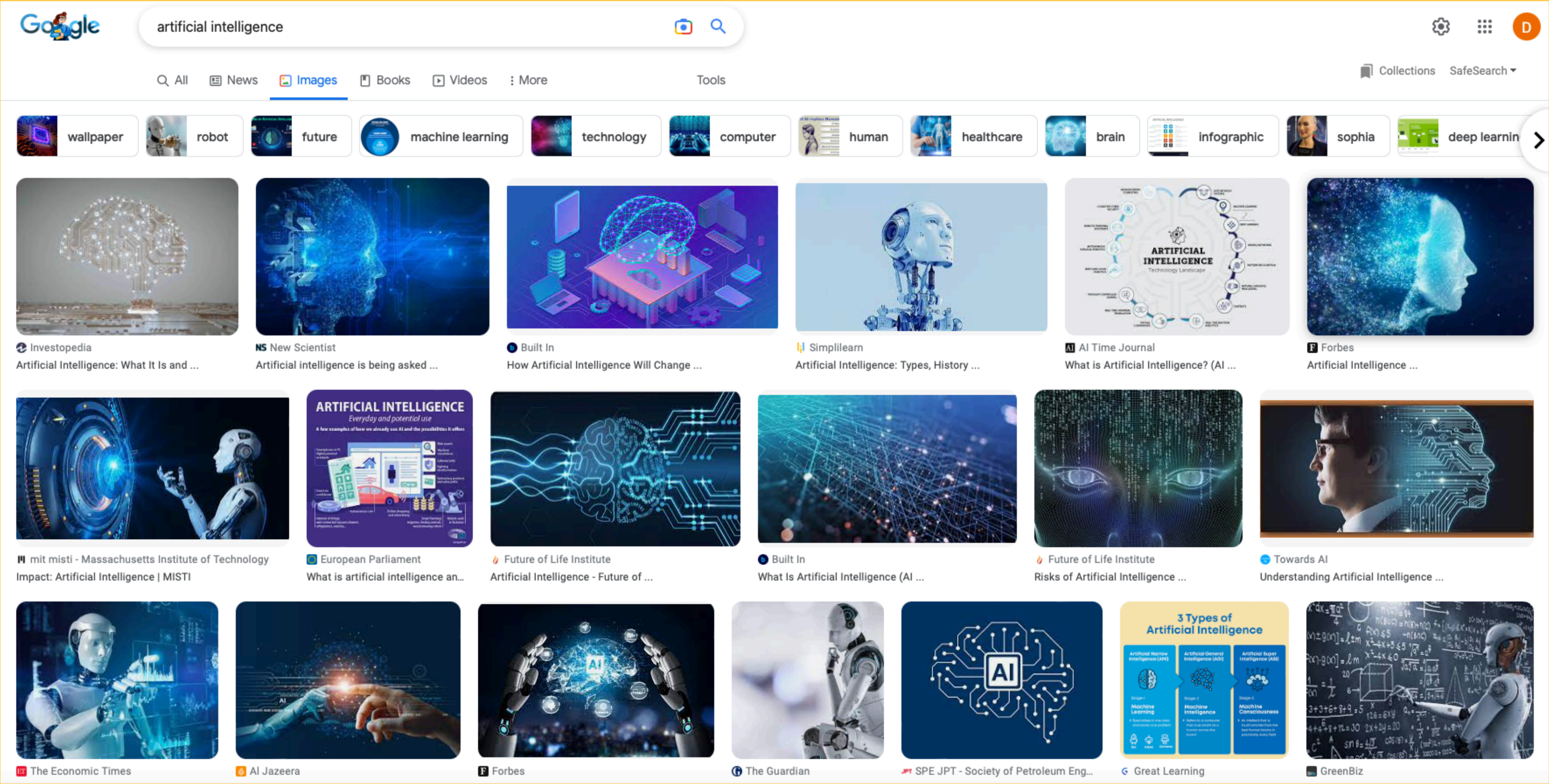
Umbrella term for a range of algorithm-based technologies that perform tasks, solve problems, and achieve objectives by carrying out functions that normally require human intelligence

AI Ethics: where to begin?

AI Ethics: where to begin?

...Let’s start with a brief discussion of what AI ethics *is not.*

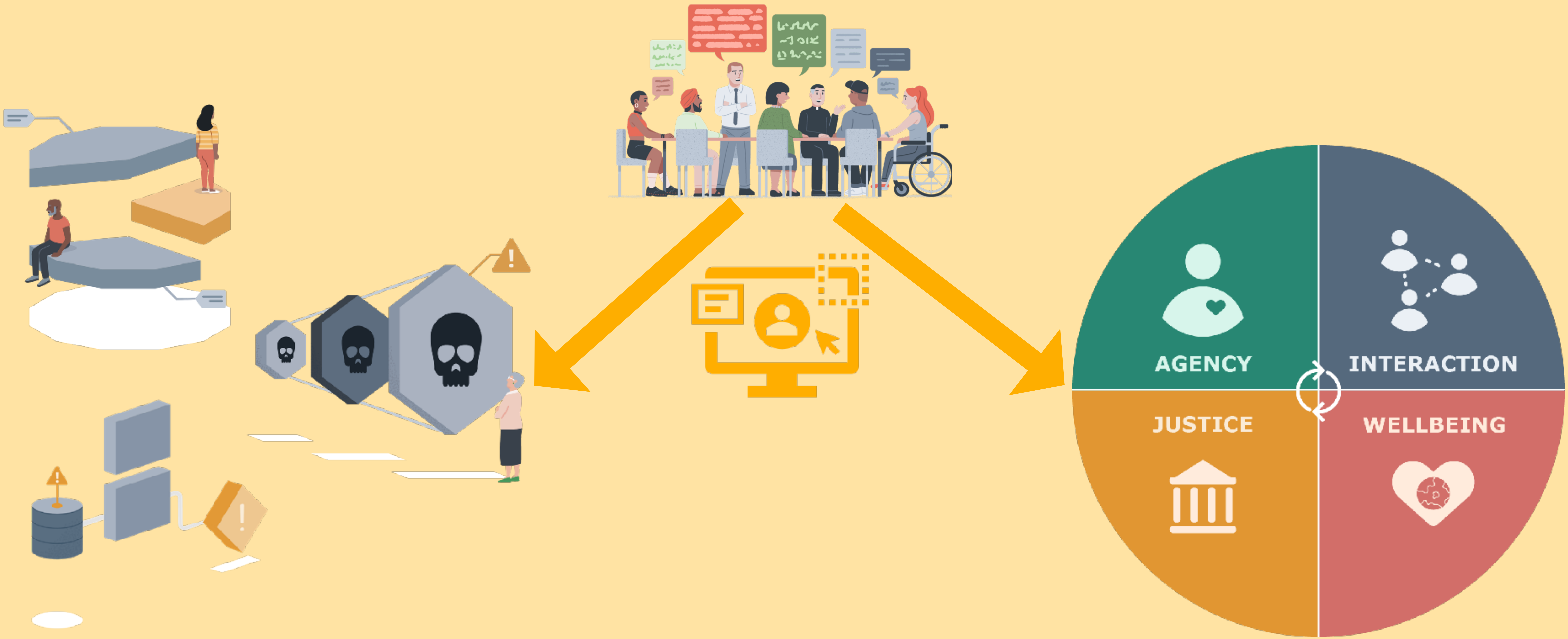
What AI Ethics *is not*



What AI Ethics *is not*

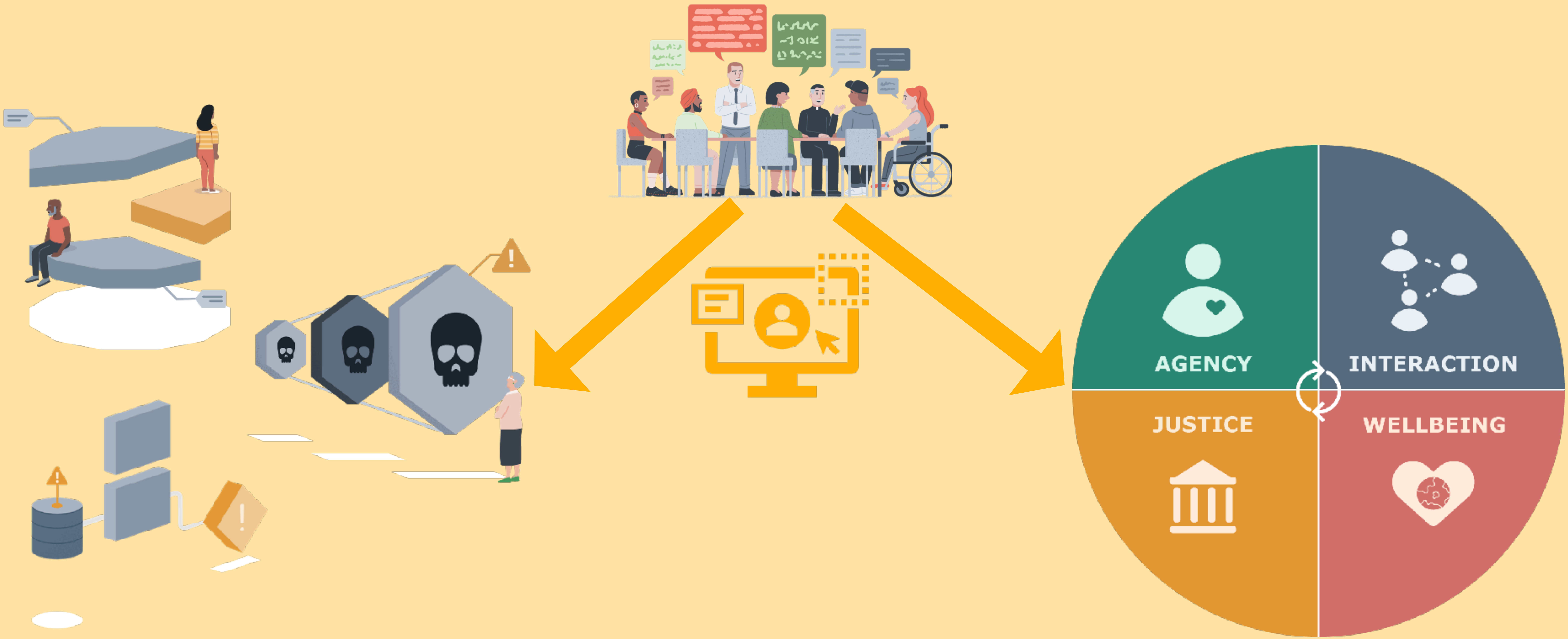
AI Ethics *is about*...

...placing all impacted humans at the axis of choice which determines the direction of technological change.



AI Ethics *is about*...

...placing all impacted humans at the axis of choice which determines the direction of technological change.



AI Ethics *is about* answering these questions

How will the values, interests, organisational cultures, and individual attitudes, standpoints, and dispositions that are currently driving the accelerating development of the AI ecosystem come to influence future society's forms of life and transform the identities of its warm-blooded subjects?

How will the technology policies and regulations that are currently born of the lagging ethical and legal vocabularies of the present keep pace with and effectively respond to the unexampled societal challenges raised by AI?

How will AI systems and the technology policies and regulations that govern them be able to steward a more sustainable and equitable cyber-physical future, in turn?

What shape will the society of tomorrow take?

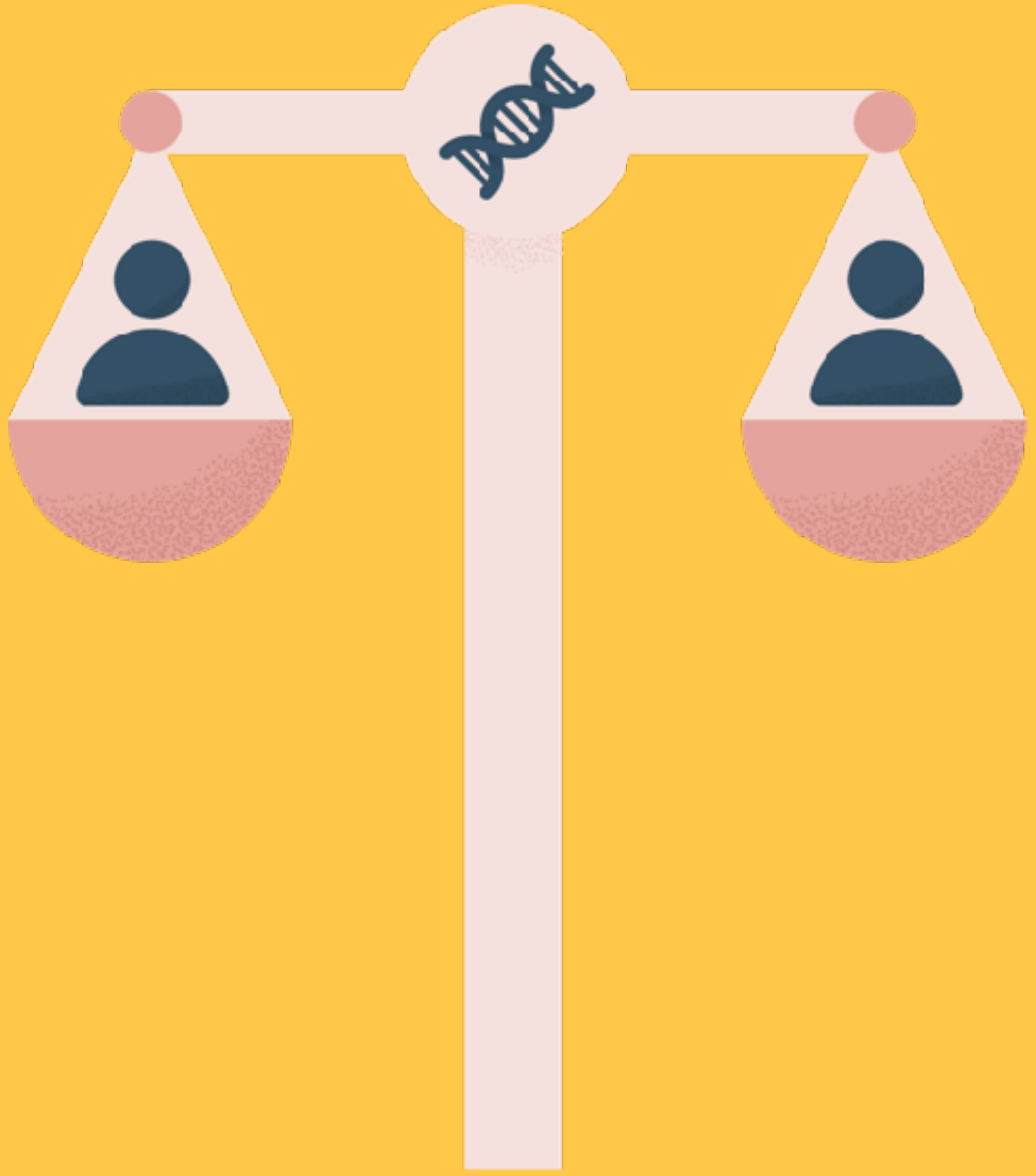
AI Ethics: So, where to begin?

Core issue when applying ethics to real-world:

Modern plurality of culture, values, and morals —> no universally accepted list of ethical values

How should AI ethicists proceed?

A more pragmatic and empirically-driven approach is needed



AI Ethics

Begins by considering the **real-world problems and harms** posed by the use of the AI and data-riven technologies

This means the scope of the values that underwrite responsible practices in AI and data-driven systems must be informed by the **actual risks posed by their use**

AI Ethics

Another valence of ethical plurality must be confronted:

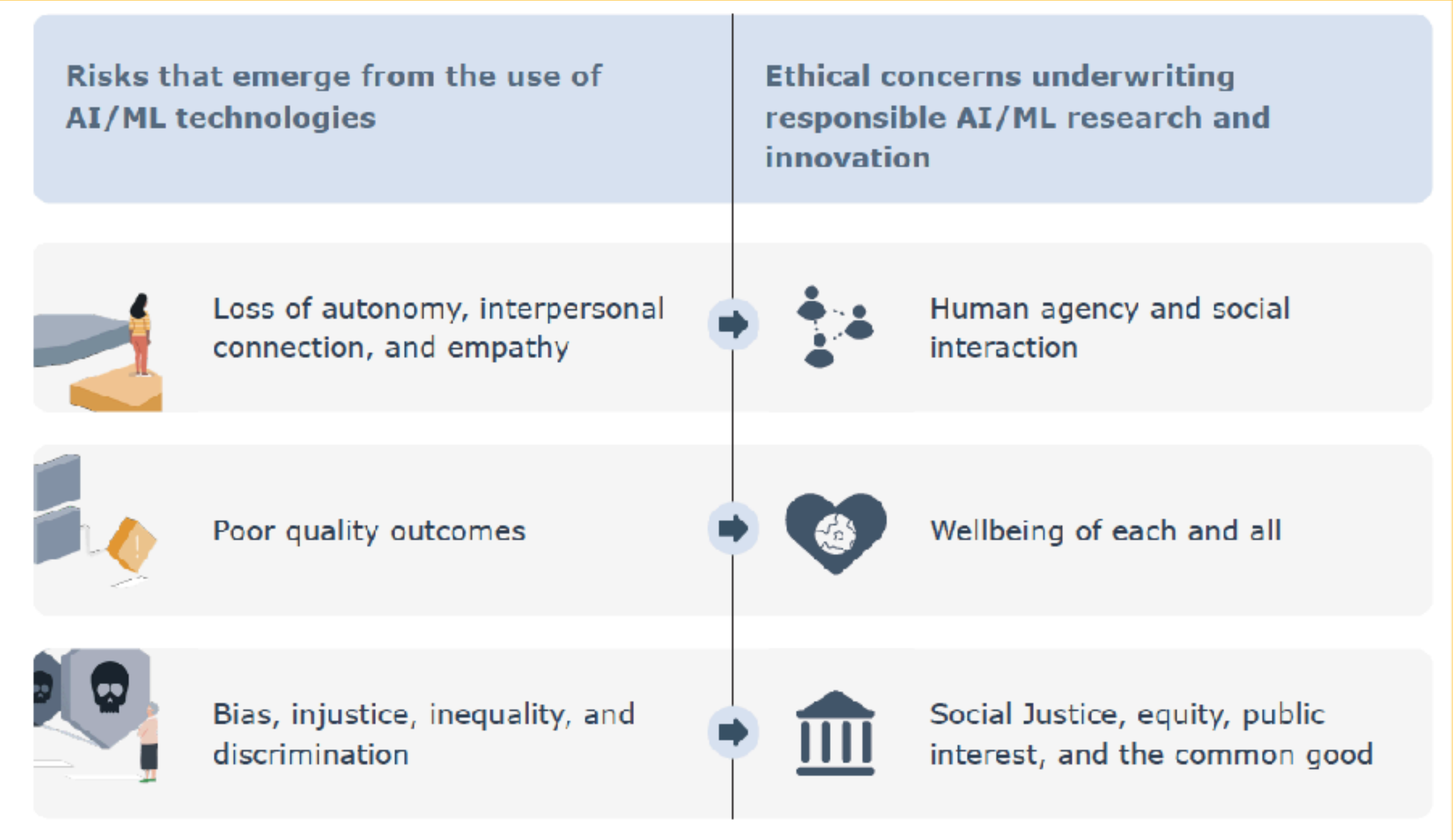
Acknowledge that the exclusion of non-Western ethical frameworks reflects deeper legacies of coloniality and Western cultural hegemony.

Redress and rectify the monoculture of Western-centric morality

Be inclusive of the diverse cultural self-understandings and lived experience of all those who may be affected

AI HARMS:

Zooming in



Risks that emerge from the use of AI/ML technologies	Ethical concerns underwriting responsible AI/ML research and innovation
Loss of autonomy, interpersonal connection, and empathy	Human agency and social interaction
Poor quality outcomes	Wellbeing of each and all
Bias, injustice, inequality, and discrimination	Social Justice, equity, public interest, and the common good

LOSS OF AUTONOMY, INTERPERSONAL CONNECTIONS, AND EMPATHY

- Automated decisions may have dehumanising effects
- Feelings of disempowerment
- Loss of sense of personal autonomy
- Loss of interpersonal connections and feelings of empathy

THREATS TO BASIC INDIVIDUAL DIGNITY, AUTONOMY, AND IDENTITY FORMATION

↗ The use of "large-scale behavioural technologies" instrumentalise targeted people and treat them as manipulable objects of prediction

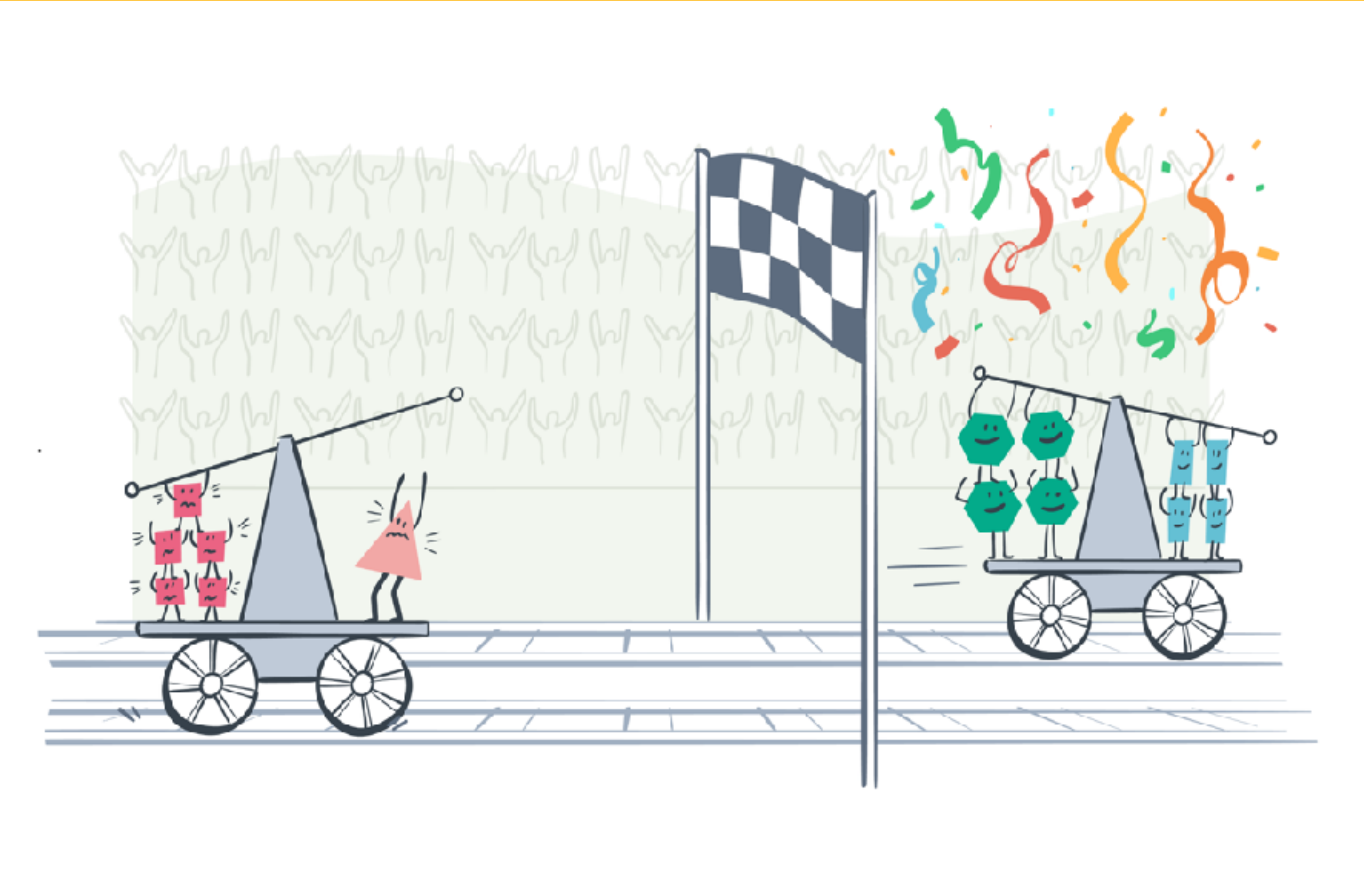
↗ Algorithmic nudging techniques promote the displacement of individual agency and the degradation of the conditions needed for the successful exercise of human judgment, moral reasoning, and practical rationality

↗ The non-consensual seizure and monopolisation of focused mental activity in the "attention market" has engendered forms of anxiety, cognitive impairment, and mental health issues

THREATS TO SOCIAL SOLIDARITY AND COMMUNICATIVE PROCESS OF SOCIAL INTEGRATION

↗ Non-consensual and opaque algorithmic production of polarised digital publics which undermine informational plurality and the deliberatively achieved political will of interacting citizens

↗ Computation-based social sorting and management infrastructures create "panoptic effects" that induce people to modify their behaviour on suspicion it is being constantly monitored and that deter open interactions which enable the development of reciprocal trust and interpersonal connection.

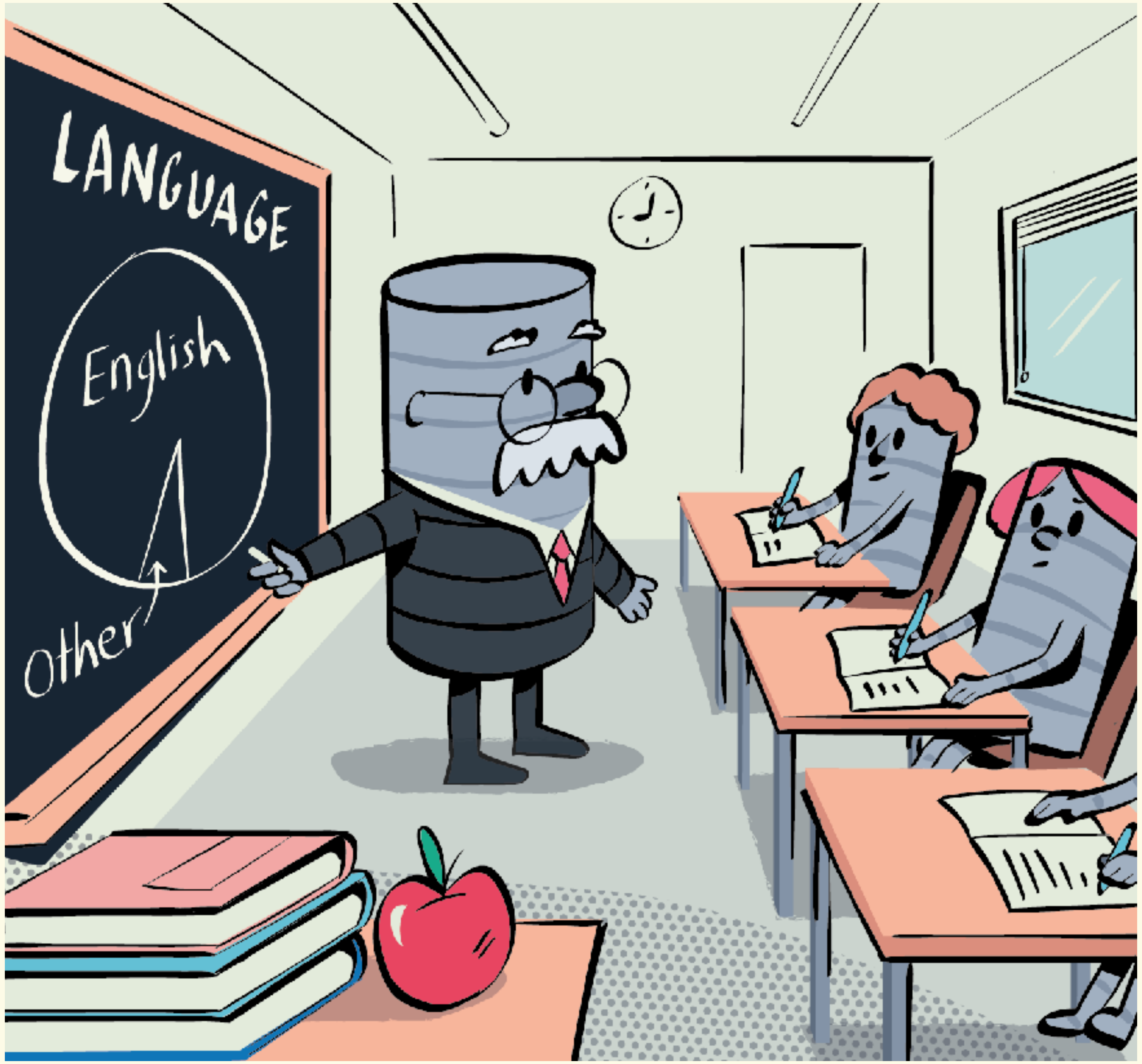


POOR QUALITY AND DANGEROUS OUTCOMES

- ‘Garbage in, garbage out’
- Measurement errors, imbalanced classes, erroneous data linkage, faulty proxies, etc.

BIAS, INEQUALITY, INJUSTICE AND DISCRIMINATION

- AI models draw insights from patterns on which they are trained.
- Model will reproduce unfair or discriminatory patterns in the data.
- AI systems learn and reproduce, sometimes even amplifying these biases or discriminatory outcomes.
- This is the case in most (if not all) datasets!



WIDENING GLOBAL AND DIGITAL DIVIDES

- Use of AI systems is not distributed equitably between countries or within regions in the same country
- Reinforcement and even amplification of the already existing data inequities and digital divides
- Data colonialism

DATA INTEGRITY, PRIVACY, AND SECURITY

- Collection, use, storage, and sharing of data can lead to multiple harms
- Contextual privacy and consent
- Concept drift, adversarial attacks, data poisoning

BIOSPHERIC HARM

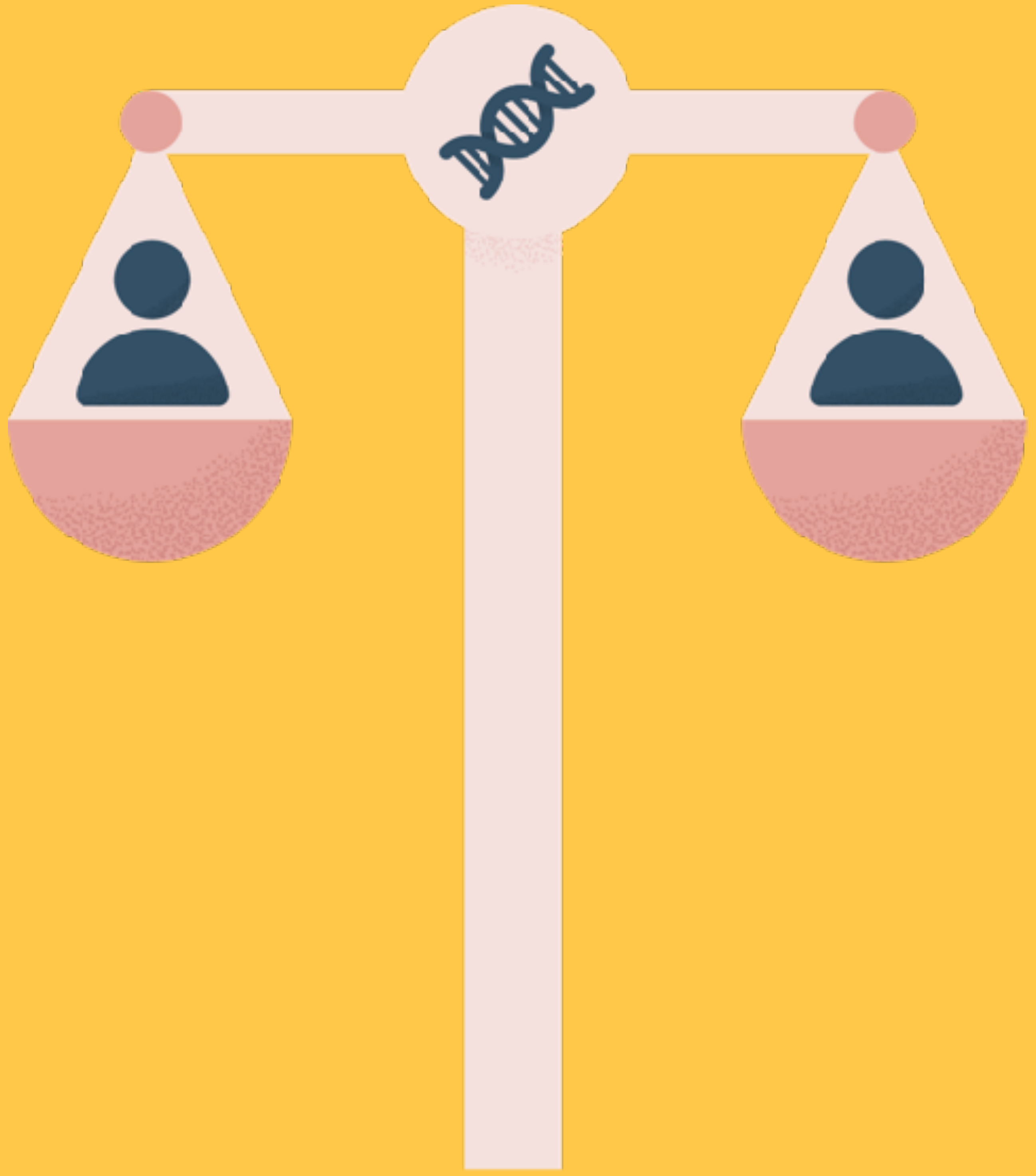
- Explosion of computing power has increasing environmental costs
- Increases in size and complexity of models translates into increasing consumption of data
- Environmental costs are not distributed equitably across society, instead follow existing patterns of environmental racism and colonialism

Questions?

Activity 1: Thinking about AI harms

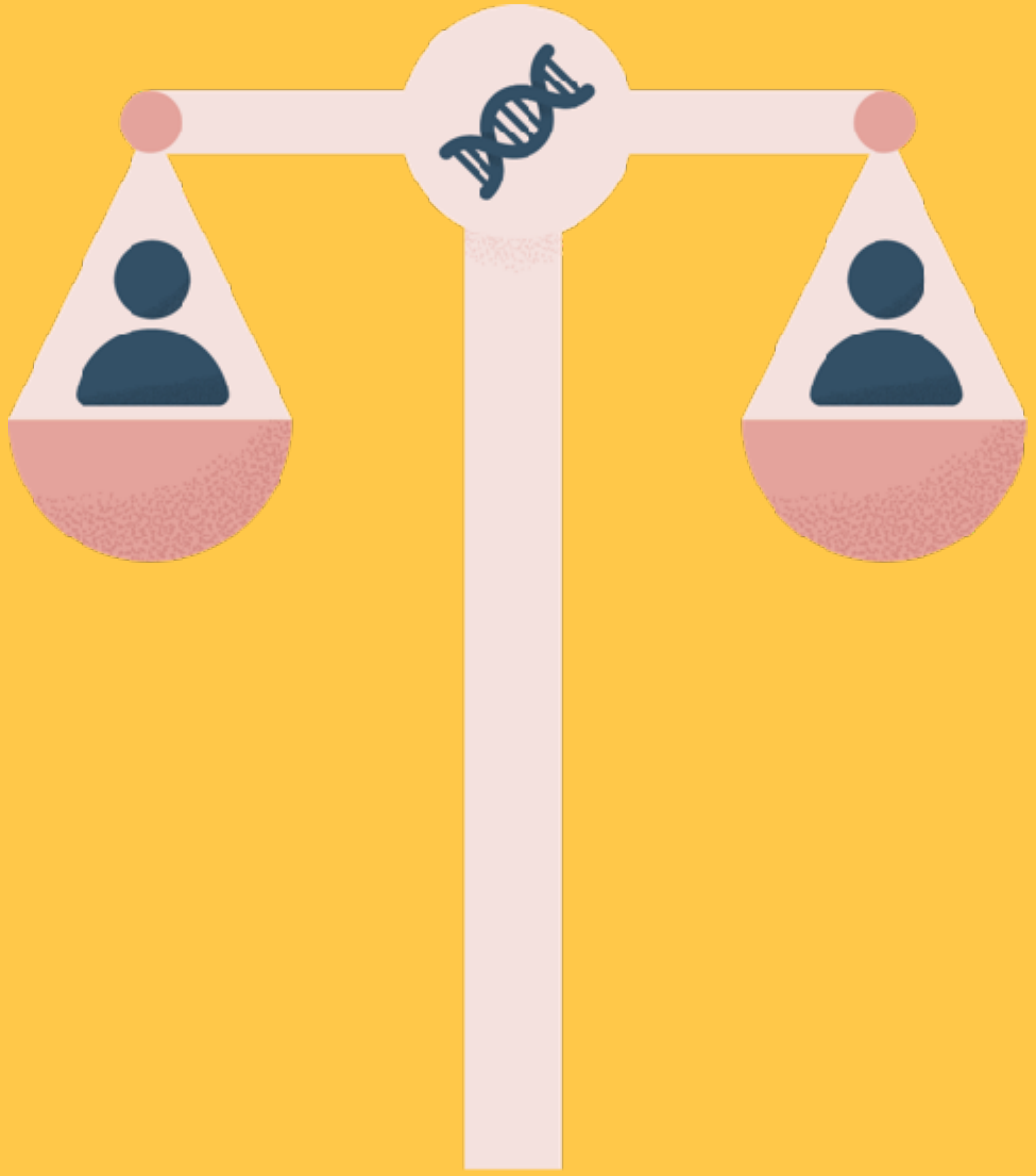
02

AI VALUES



Context: bioethics and human rights

The appeal of bioethics and human rights as starting point for AI Ethics is rooted in in the dynamics of real-world harm



Context: bioethics and human rights

Human rights are the basic rights and freedoms that are possessed by every person in the world and that preserve and protect the inviolable dignity of each individual. These create obligations that bind governments to respecting, protecting, and fulfilling human rights.

The main principles of bioethics include respecting the autonomy of the individual, protecting people from harm, looking after the well-being of others, and treating all individuals equitably and justly.

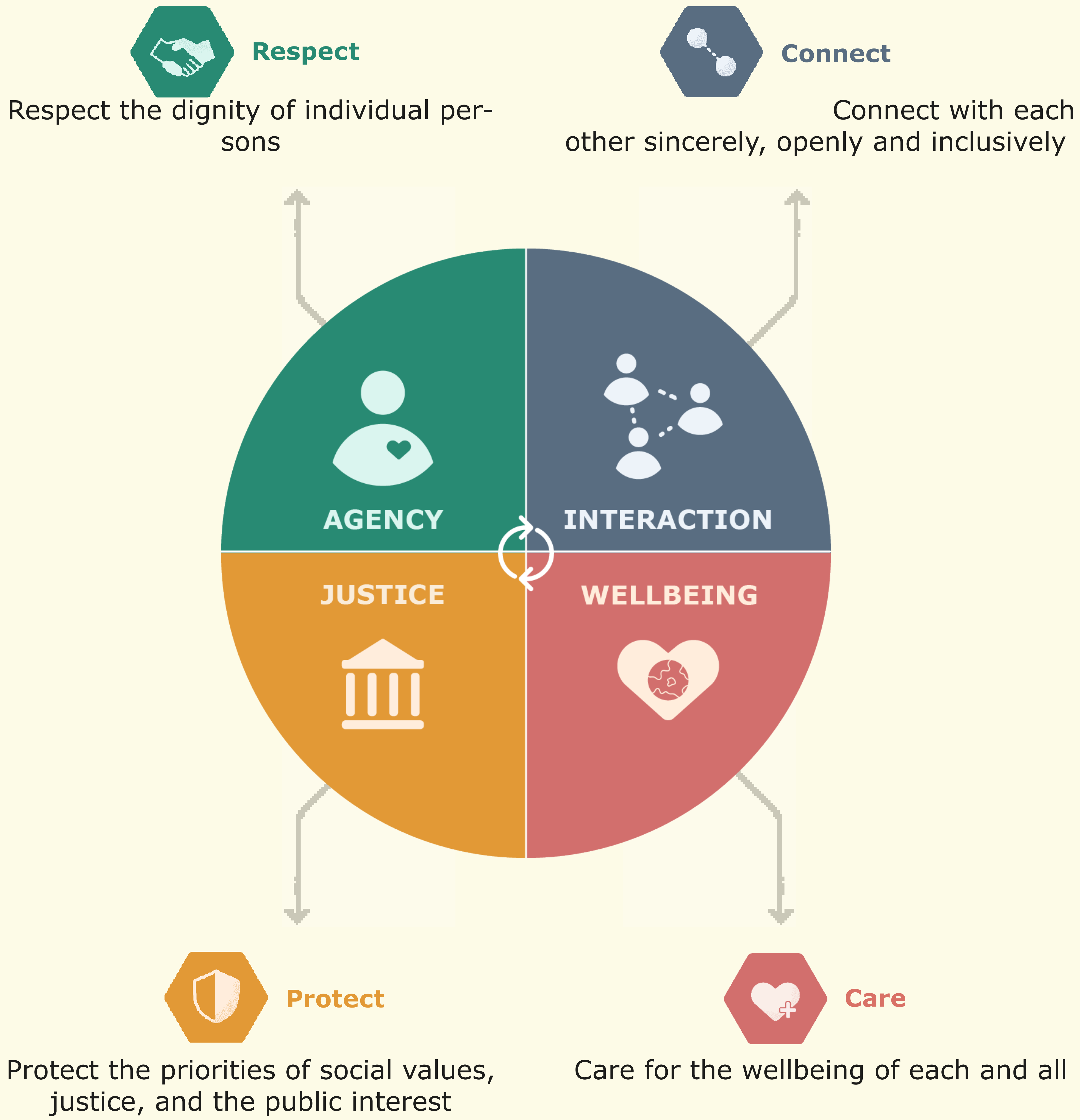
SUM Values and GPAI Principles

Support

Underwrite

Motivate

… a responsible AI innovation system


Respect
Respect the dignity of individual persons
Connect
Connect with each other sincerely, openly and inclusively
AGENCY
INTERACTION
JUSTICE
WELLBEING
Protect
Protect the priorities of social values, justice, and the public interest
Care
Care for the wellbeing of each and all

RESPECT

... the dignity of individual persons

1/ Ensure people's abilities to make free and informed decisions about their own lives

2/ Safeguard their autonomy, their power to express themselves, and their right to be heard

3/ Secure their capacities to make well-considered and independent contributions to the life of the community

4/ Value the uniqueness of their aspirations, cultures, contexts, and forms of life

5/ Secure their ability to lead a private life and to flourish, to fully develop themselves

RESPECT

Ethical Concerns

- Dignity, autonomy, agency, and authority of persons
- Self-realisation and flourishing of individuals

Related human rights and fundamental freedoms

- The right to human dignity
- The right to life
- The right to liberty and security
- Freedoms of thought, conscience, and religion
- Freedom of expression and opinion
- The right to respect for private and family life and the protection of personal data

CONNECT

... with each other sincerely, openly and inclusively

1/ Safeguard the integrity of interpersonal dialogue, meaningful human connection, and social cohesion

2/ Prioritise and encourage diversity, participation, inclusion, and consideration of all voices

3/ Encourage all voices to be heard and all opinions to be weighed seriously and sincerely

4/ Utilise AI innovations pro-socially to enable bonds of interpersonal solidarity to form and to foster the capacity to connect, to reinforce trust, empathy, and reciprocal responsibility

CONNECT

Ethical Concerns

- Integrity of interpersonal relationships,
- Solidarity
- Participation-based innovation and stakeholder inclusion

Related human rights and fundamental freedoms

- Freedom of assembly and association
- The right to diverse and reliable information and access to plurality of ideas and perspectives.
- The right to participate in the conduct of public affairs and good governance

CARE

... for the wellbeing of each and all

1/ Design and deploy AI systems to foster the welfare of all stakeholders

2/ Do no harm with these technologies and minimise the risks of their misuse

3/ Prioritise the safety and the mental and physical integrity of people when scanning horizons of technological possibility and when conceiving of and deploying AI applications

CARE

Ethical Concerns

- Beneficence, safety, and non-harm
- Stewardship of individual, communal and biospheric wellbeing

Related human rights and fundamental freedoms

- The right to life
- The right to physical, mental, and moral integrity.
- Environmental sustainability as foundation for the enjoyment of all rights and freedoms

PROTECT

... the priorities of social values, justice, and the public interest

1/ Use digital technologies as an essential support for the protection of fair and equal treatment under the law

2/ Prioritise social welfare, public interest, and the social and ethical impacts of innovation in determining the legitimacy and desirability of AI technologies

3/ Use AI to empower and to advance the interests and well-being of as many individuals as possible

4/ Think big-picture about the wider impacts of the AI technologies you are conceiving and developing. Think about the ramifications of their effects and externalities for others around the globe, for future generations, and for the biosphere as a whole

PROTECT

Ethical Concerns

- Justice and equity
- Prioritisation of the public interest and common good

Related human rights and fundamental freedoms

- Prohibition of discrimination and the right to non-discrimination
- Equality before the law
- The right to an effective remedy for violation of rights and freedoms
- The right to a fair trial and due process
- The right to judicial independence and impartiality
- Equality of arms

INTERCONNECTIVITY, SOLIDARITY, AND INTERGENERATIONAL RECIPROCITY

Value	Elaboration & Corresponding Rights and Freedoms	Resources for Principles and Priorities and Corresponding Rights and Freedoms
Interconnectivity, solidarity, and intergenerational reciprocity	*All humans are interconnected to a greater whole, which transcends time and thrives when all its constituent parts are enabled to thrive. This unbounded bond of solidarity extends from the closest relationship between kin to the living totality of the biospheric whole. Membership in this greater community also places a responsibility on the present generation to take account of the well-being and flourishing of future generations. Intergenerational reciprocity involves looking backward in considering the wisdom and learning of past generations and looking forward in considering the rights and well-being of lives not yet lived (two, four, seven, or more generations in the future).* ~ **-The right of future generations to due moral regard and consideration** **- *Kaitiakitanga* (Maori): The responsibility to ensure sustainable futures for the biosphere and for people, families, communities, and humanity** **- *Manaakitanga* (Maori): The responsibility to extend care, compassion, hospitality, and generosity to all others including strangers and the environment. Shared *Manaakitanga* supports well-being, dignity, and the stewardship of healthful and spiritual living.** **-The Seventh Generation Principle (Haudenosaunee Confederacy, Iroquois): Give regard to the well-being of the seventh generation ahead of you in your practices, works, actions, and deliberations and draw on the experience and wisdom of the seventh generation that came before** **-The values of *Ubuntu* (Sub-Saharan Africa): Ethical life is measured by the meaningful relationships formed by each individual with an interconnected and interdependent whole of people, community, and environment. One's humanity is affirmed by connecting with and taking care of others and by recognizing their dignity in works, deliberations, and deeds.**	**UNESCO:** -III.1 Values, Recommendation on the Ethics of Artificial Intelligence, *Living in peaceful, just and interconnected societies* **Other resources:** The Maori Report, **Independent Maori Statutory Body** Treaty of Waitangi/Te Tiriti and Māori Ethics Guidelines for: AI, Algorithms, Data and IOT, **2020** The World People's Conference on Climate Change and the Rights of Mother Earth, **Bolivia 2010** The Constitution of the Iroquois Nations, **1916** What is Ubuntu?, **Desmond Tutu 2013** I am because you are, **Michael Onyebuchi Eze, UNESCO 2011**

ENVIRONMENTAL FLOURISHING, SUSTAINABILITY, AND THE RIGHTS OF THE BIOSPHERE

Value	Elaboration & Corresponding Rights and Freedoms	Resources for Principles and Priorities and Corresponding Rights and Freedoms
Environmental flourishing, sustainability, and the rights of the biosphere	*All humans draw oxygen from the Earth's air, draw nourishment from its soil, and live as interconnected parts of a living biospheric community. The interrelated organisms of this unbounded community share a common origin, a common history, and a common ecological fate. Members of humanity, as benefactors and inheritors of such a circle of life and of the life-giving gifts of the earth, should seek practices of living that secure environmental flourishing, sustainability, and the rights of the biosphere. These practices of living should aim for a harmony and balance with the interdependent ecologies of the biosphere in solidarity with it. They should also respect nature's right to flourish, to endure, and to regenerate life without harmful anthropogenic influence. All people involved in AI and data innovation lifecycles should prioritise environmental flourishing, sustainability, and the rights of the biosphere, ensuring that they use the affordances of technology to do battle with climate change and biodiversity drain rather than contribute to them.* ~ **-The right of *Pachamama*: 'Nature or *Pachamama*, where life is reproduced and exists, has the right to exist, persist, maintain and regenerate its vital cycles, structure, functions and its processes of evolution'. (Article 1, Constitution of Ecuador)** **-*Sumak kawsay* (Quechua), *suma qamaña* (Aymara), *buen vivir* (Spanish): "living well" or "collective well-being" but also the priority of a shared pursuit of the fullness, creativity, harmony, and flourishing of human and biospheric life.** **- *Kaitiakitanga* (Maori): The responsibility to ensure sustainable futures for the biosphere and for people, families, communities, and humanity** **- 'Environmental Justice affirms the sacredness of Mother Earth, ecological unity and the interdependence of all species, and the right to be free from ecological destruction'. (First National People of Colour Environmental Leadership Summit)**	**UNESCO:** -III.1 Values, Recommendation on the Ethics of Artificial Intelligence, *Environment and ecosystem flourishing* **Other resources:** The Constitution of Ecuador, **2008** 17 Principles of Environmental Justice, **First National People of Colour Environmental Leadership Summit 1991** Bali Principles of Climate Justice, **2002** The Maori Report, **Independent Maori Statutory Body** Treaty of Waitangi/Te Tiriti and Māori Ethics Guidelines for: AI, Algorithms, Data and IOT, **2020** The World People's Conference on Climate Change and the Rights of Mother Earth, **Bolivia 2010** The Albuquerque Declaration, **Native People-Native Homelands Climate Change Workshop-Summit, Albuquerque, New Mexico, 1998**

Any questions?

Activity 2: Relating to values

Go to link [insert link for MIRO Board]


Thank you!
See you tomorrow for Day 3: AI Sustainability