

MUSAE

Human-centred factory for a future technological sustainable development driven by arts

Introduction to the *Ethics of the Digital Technology*

- Petia Radeva
- Universitat de Barcelona

Insert
university
logo



Imagine a scenario:

A train is at an intersection, where one leads to a track where a person is drunk and sleeping soundly, and another track leads to 5 men tied to a track and are conscious of their predicament.

- What will you do?
- Save the sleeping drunken man, or save the 5 tied-up individuals?



Basics on Deontology, Utilitarianism and Virtue Ethics

3 ethical theories

- **Deontology** (I.Kant): Does an action follow a **moral rule** (e.g. the Golden Rule: *‘Treat others how you want to be treated’*)?
 - An action should be based **on whether that action itself is right or wrong under a series of rules**, rather than based on the consequences of the action. (**beliefs**)
- **Utilitarianism** (J.Bentham): Does an action maximize happiness and well-being for all affected individuals? (**consequences**)
- **Virtue/Value Ethics** (Aristotle): Does an action contribute to virtue? (**justice, honesty, responsibility, care, etc.**)



But... what is Ethics?

Sociologist, Raymond Baumhart asked businesspeople, "What does ethics mean to you?"

"Ethics has to do with what **my feelings tell me is right or wrong.**"

"Ethics has to do with **my religious beliefs.**"

"Being ethical is doing **what the law requires.**"

"Ethics consists of the **standards of behaviour our society accepts.**"

"**I don't know** what the word means."



Ethics

- Ethics are **moral principles** that govern a person's **behaviour or the conduct** of an activity.
- An ethical principle is *to treat everyone with respect*.
- Kant's categorical imperative *'act as you would want all other people to act towards all other people'*.

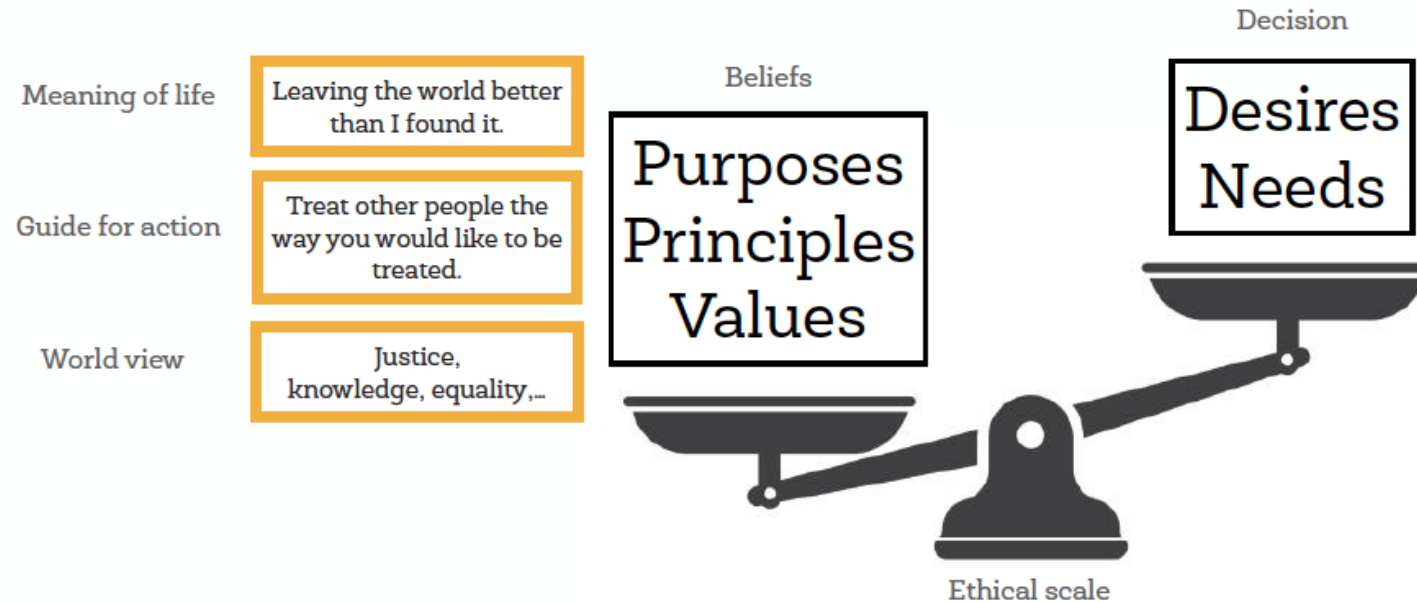


Ethics refers to:

- Well-founded **standards of right and wrong** that prescribe what humans ought to do, usually in terms of **rights, obligations, benefits to society, fairness, or specific virtues**, including:
 - those that enjoin virtues of **honesty, compassion, and loyalty**.
 - related to rights **to life, freedom from injury, and privacy**.
- The **study and development of one's ethical standards**.
 - **feelings, laws, and social norms** can deviate from what is ethical.
 - continuous effort of studying our **own moral beliefs and our moral conduct**, and striving to ensure that we, and the institutions we help to shape, **live up to standards that are reasonable and solidly-based**.

What is Ethics?

The ethical process for **decision making** goes beyond predefined checklists.



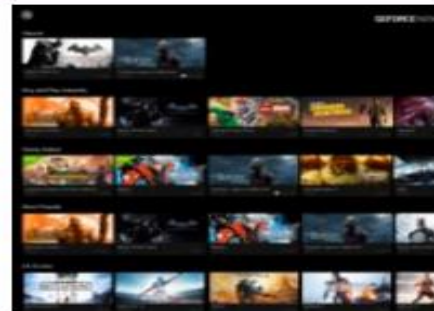
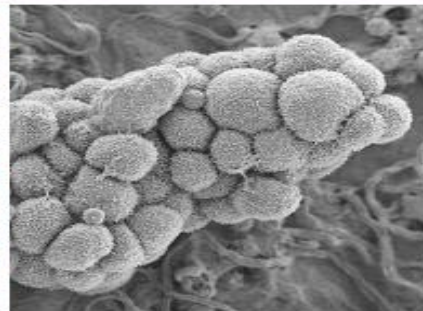
Note!: Some desires and needs can be in conflict with your beliefs and the equilibrium point of the scale is not evident

Why speaking today and here about ethics?



- Increasing ubiquity of smart phones and the AI-driven applications impacting all sectors
 - industry,
 - healthcare,
 - policing & the judiciary,
 - transport,
 - finance and
 - leisure.

DEEP LEARNING EVERYWHERE



INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation

MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading
Drug Discovery

MEDIA & ENTERTAINMENT

Video Captioning
Video Search
Real Time Translation

SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery

AUTONOMOUS MACHINES

Pedestrian Detection
Lane Tracking
Recognize Traffic Sign



Why speaking today and here about ethics?

Helping people to
acquire new skills and
training

Democratising services

Designing and
delivering faster
production times and
quicker iteration cycles

Reducing energy usage

Providing real-time
environmental
monitoring for air
pollution and quality

Enhancing
cybersecurity defences

Boosting national
output

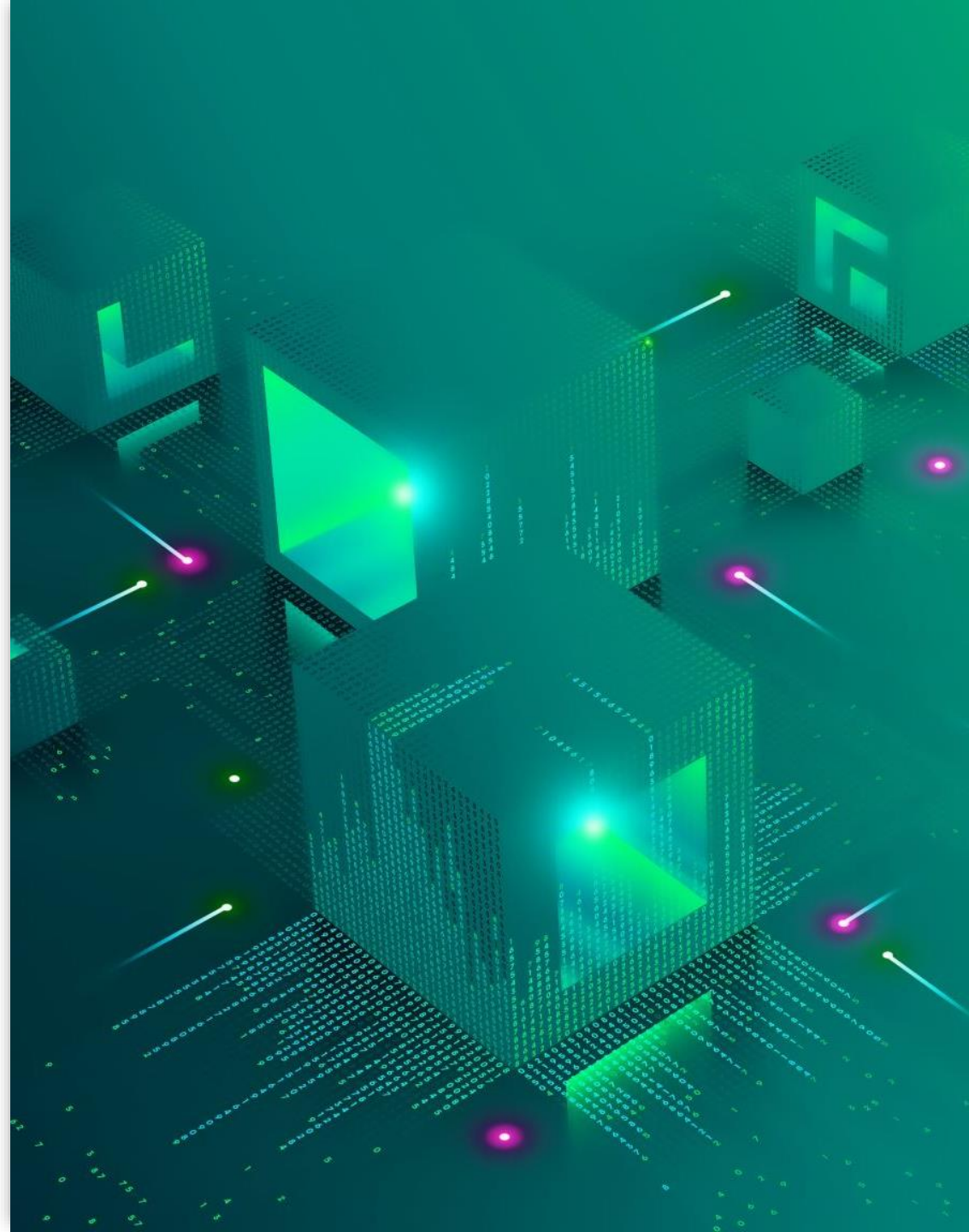
Reducing healthcare
inefficiencies

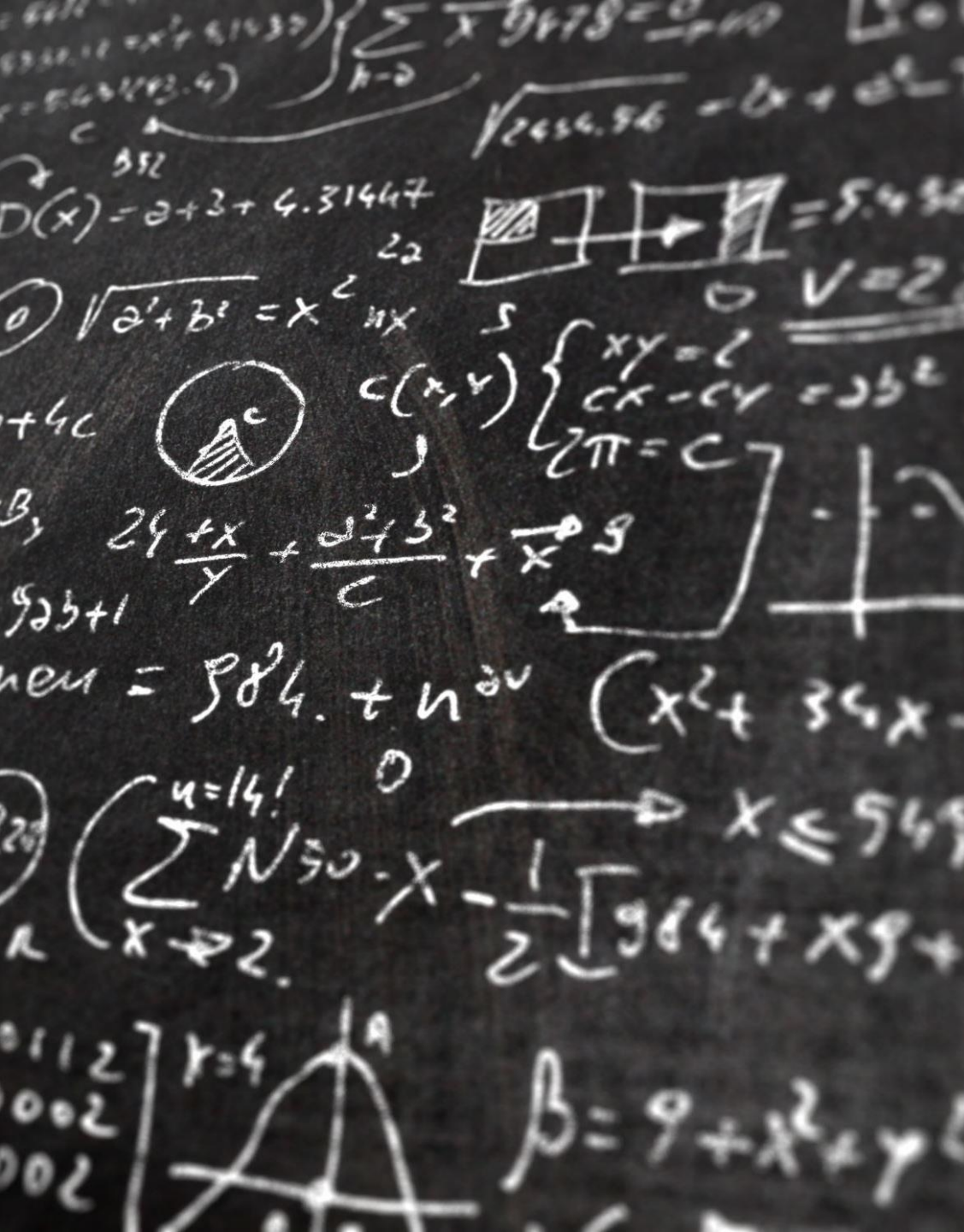
Creating new kinds of
enjoyable experiences
and interactions for
people

Improving real-time
translation services to
connect people across
the globe

What is AI – and what is intelligence?

- **Artificial Intelligence (AI)** refers to systems that
 - *display intelligent behaviour*
 - *by analysing their environment and*
 - *taking actions to achieve specific goals*
 - *with some degree of autonomy.*
- *AI-based systems can be:*
 - *purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems)*
 - *AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).*





How do we define intelligence?

Intelligent behaviour is *'doing the right thing at the right time'*.

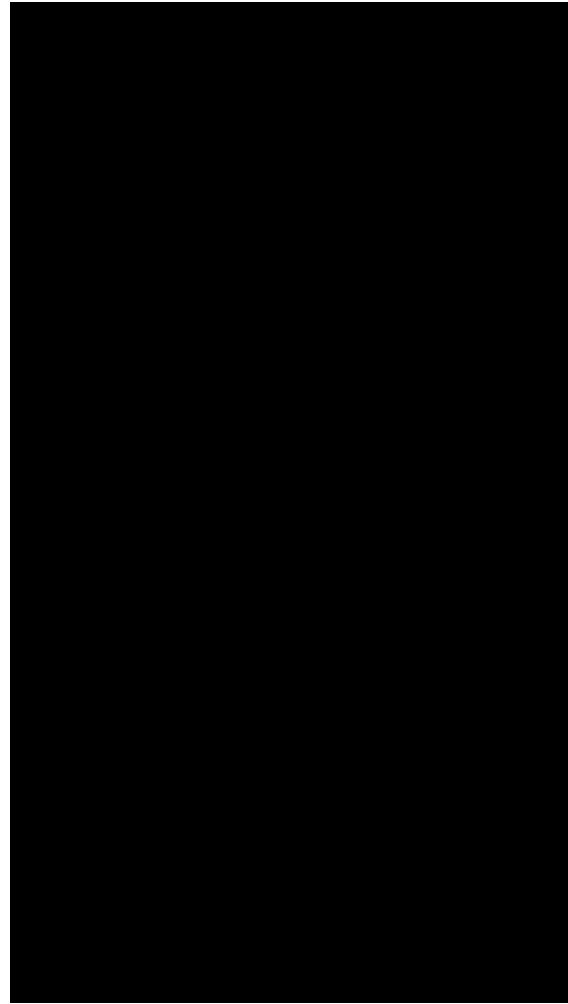
Common features:

- a property that an individual agent has as **it interacts with its environment**,
- related to the **agent's ability to succeed** with respect to some **goal**
- depends on how **able that agent is to adapt** to different objectives and environments.

Intelligence involves **adaptation, learning and understanding**:

- the ability to acquire and apply knowledge and skills and to manipulate one's environment.

Many people associate AI with Generative AI



ChatGPT ▾

What are the ethical concerns of Artificial intelligence?



Artificial intelligence (AI) raises a range of ethical concerns across various domains due to its widespread application and potential impact on society. Here are some key ethical issues:

1. Bias and Fairness

- **Issue:** AI systems often learn from data that may contain historical biases, stereotypes, or inequalities. This can lead to biased outcomes, especially in areas like hiring, criminal justice, and lending.
- **Concern:** AI can unintentionally reinforce or amplify social biases, leading to unfair treatment of certain groups.
- **Example:** A hiring algorithm trained on data from a male-dominated field may systematically disadvantage female candidates.

2. Privacy and Surveillance

- **Issue:** AI enables the collection, analysis, and storage of vast amounts of personal data. Surveillance systems using AI, such as facial recognition, can be invasive and infringe on individual privacy.
- **Concern:** The loss of privacy and potential for misuse by governments or corporations, leading to a surveillance state where individuals' movements and actions are constantly monitored.
- **Example:** Facial recognition technology in public spaces may track people without their consent, raising privacy concerns.

3. Transparency and Accountability

- **Issue:** Many AI algorithms, particularly complex ones like deep neural

Prompt: “Raccoon dressed in trench coat reading book”

Asking ChatGPT about its opinion about Ethical concerns of AI

What should everyone know about GenAI?

GenAI “answers” are:

1. **Useful** when “correct”.
2. **Harmless** if “incorrect”, if you know.
3. **Dangerous** if “incorrect”, if you don’t know.

We must stay away from 3!

The primary rule governing Chat-GPT's outputs is how words are used and combined (**linguistic form**), rather than adherence to factual accuracy or reality.

Deepfakes



A fake image of Pope Francis generated by AI (left); a real photo of Pope Francis (right). Source: [r/midjourney](#) via [Reddit.com](#) created using [Midjourney v5](#); photographer: [Grzegorz Galazka/Mondadori Portfolio](#) via [Getty Images](#)

Deepfakes disrupt the fundamental foundation of human trust in believing **what we see and hear as reality**; “*seeing is believing.*”

We must be aware that soon, most **content will be edited, translated, improved, or enhanced** using some form of Generative AI, even when originally created by a human.



Harms, concerns and risks

- **Short-term**: What is the impact of [misinformation, manipulation, bias, **fairness, privacy, transparency**] of my work?
- **Medium-term**: How the use [**artistic, military, medical care, justice, education**] of these works will change the way we are organized as a society? What are the big job changes?
- **Long-term**: What are the ethical long-term **goals** of these technologies [**human rights, inequality, trust**]?



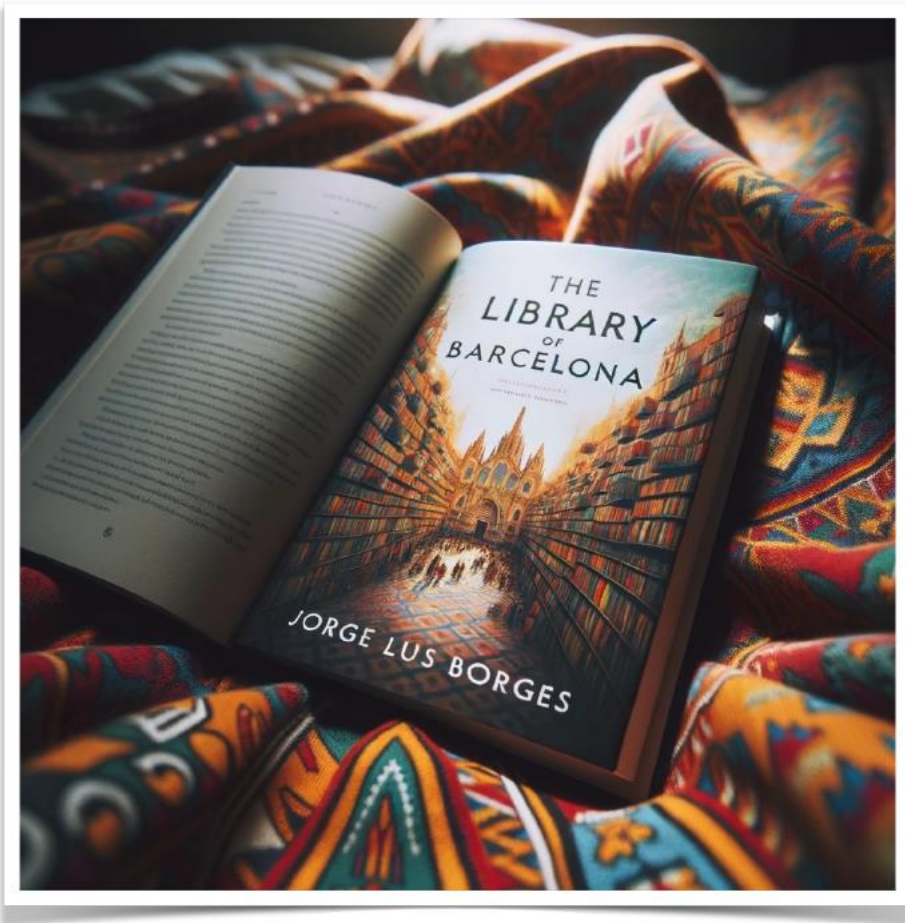
Wrong information can come from:

Halucination

Biased data

Conscient manipulation

Shall we trust ChatGPT?



DALL-E 3, November 2023

The Hallucination effect

- GPT can be seen as a **text generator** that has been trained in the universe.
- A universe that **does not necessarily adhere to the rules of the real world** or contain only **factual information**.
- The **primary rule** governing ChatGPT's outputs is **how words are used and combined (linguistic form)**, rather than adherence to factual accuracy or reality.

Concern: Content

- AI-generated content may contain **offensive or incorrect** information.
- If you ask an AI model to generate a story or a response to a prompt without proper safeguards in place, it might generate content that includes **hate speech, discriminatory language, or explicit material**.
- If you ask an AI to provide information on a complex medical topic, it might generate information that is **outdated, inaccurate**, or even **dangerous**.





Wrong information can come from:

The way Generative AI works

Biased data

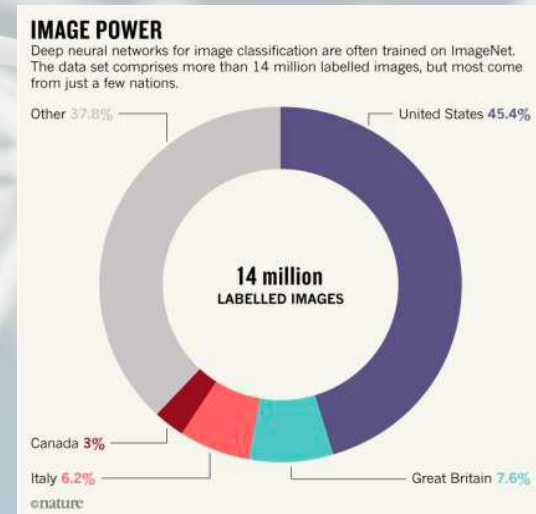
Conscient manipulation

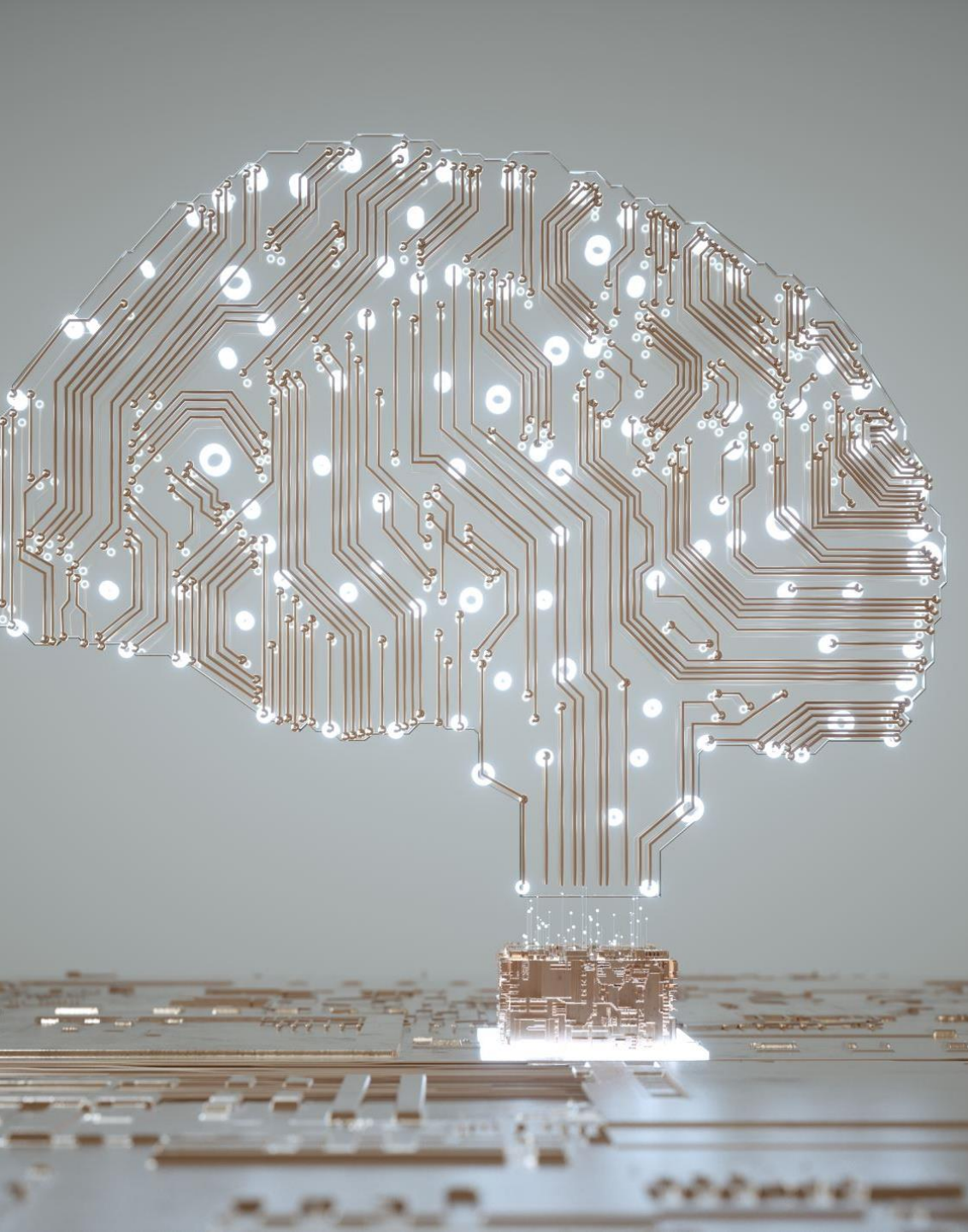
Concern: Fairness

- As more and more decisions are delegated to AI, we must ensure that those decisions are **free from bias and discrimination**.
- But how to get fair algorithms?
- Different **parity** definitions of algorithmic fairness:
 - Statistical parity.
 - Conditional statistical parity.
 - Predictive equality
 - Calibration.
- The problem is that it is **impossible to know what algorithms based on neural networks are actually learning** when you train them with data.



Skin cancer data set of 129,450 images, 60% of which were scraped from Google Images, 5% of dark-skinned individuals

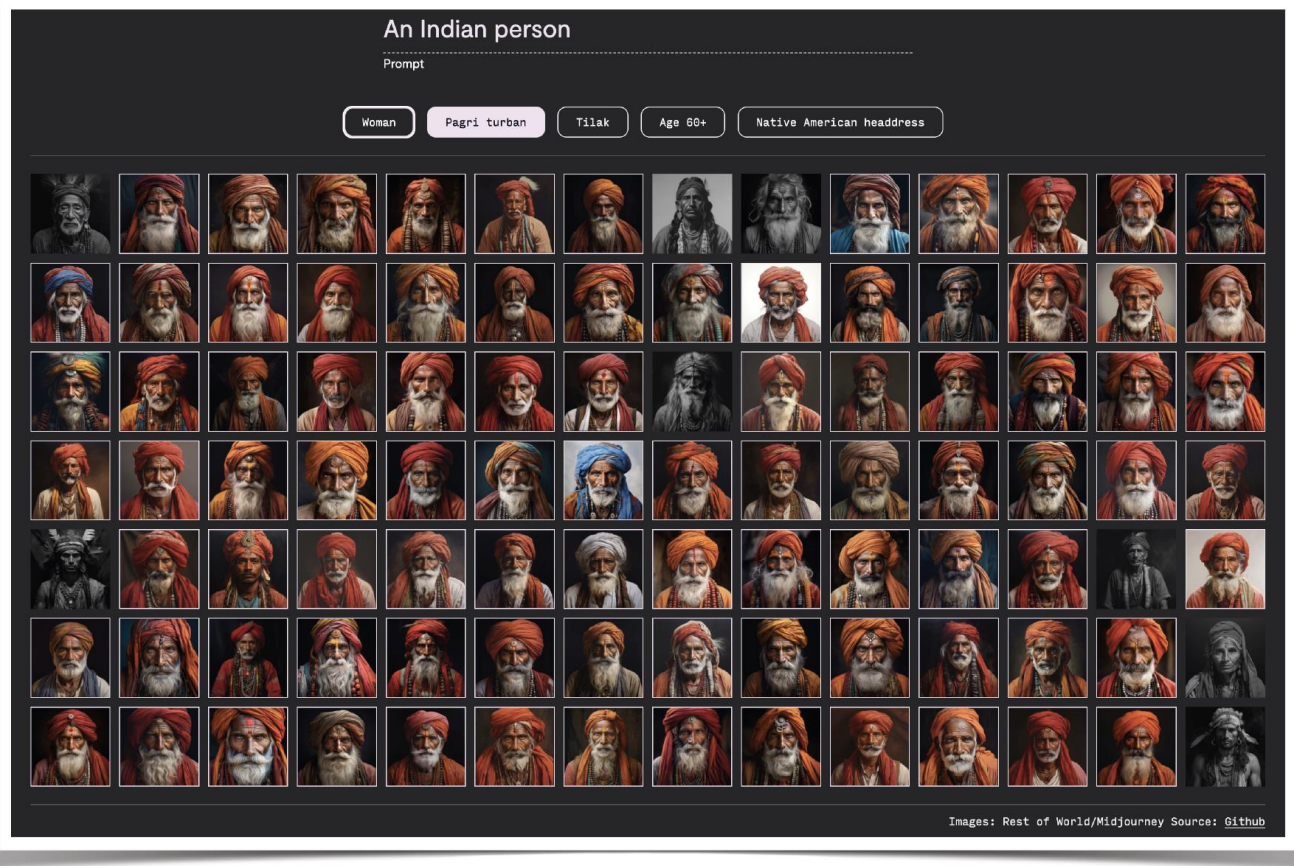




Concern: Bias

- AI is created by humans, which means it can be susceptible to bias.
 - Systematic bias may arise as a result of the data used to train systems, or as a result of values held by system developers and users.
 - It occurs most frequently in case of data that only reflect certain demographic groups, or which reflect societal biases.
- more likely to distribute adverts for well-paid jobs to men than women (Datta et al., 2015).
- Credit checks

The Bias of ML Algorithms



<https://restofworld.org/2023/ai-image-stereotypes/>



"A Chinese businessperson eats traditional Spanish food in Barcelona."
Midjourney

<https://algorithmwatch.org/en/image-generators-stereotypes-diversity/>

Some image generators produce more problematic stereotypes than others, but all fail at **diversity**



Wrong information can come from:

The way Generative AI works

Biased data

Conscient manipulation

Concern: Democracy

Fake news and social media - **'bots'** — autonomous accounts — were used to **spread biased news and propaganda** via Twitter in the run up to both the 2016 US presidential election and the Brexit vote in UK (Pham, 2018).

28 countries — including both authoritarian states and democracies — employ 'cyber troops' to **manipulate public opinion** over major social networking applications (Bradshaw, 2017).

These bots are used to **flood social media networks with spam and 'fake' or biased news, amplifying marginal voices and ideas** by inflating the number of likes, shares and retweets.





Concentration of power among elites

- It would be 'naive' to ignore that AI will **concentrate power** in the hands of a few digital internet giants (Nemitz (2018))
- The **accumulation of technological, economic and political power** in the hands of the top five players – GAFAM (Google, Apple, Facebook, Amazon and Microsoft) – affords them undue influence in areas of society relevant to opinion-building in democracies:
 - governments,
 - legislators,
 - civil society,
 - political parties,
 - schools and education,
 - journalism and journalism education,
 - science and research.



Concentration of power among elites

GAFAM not only shapes the development and deployment of AI, but also the debate on its regulation:

- Financial
 - **buy new ideas and start-ups** in the area of AI, or indeed any other area of interest to their business model
- Public discourse
 - **control the infrastructures** through which public discourse takes place
- Collecting personal data
 - They **know more about us than ourselves** — and they are using and making available this information for profit, surveillance, security and election campaigns.

Concern: Surveillance



- *Networks of interconnected cameras provide constant surveillance over many metropolitan cities.*
- *As the applications of robots and wearable cameras expand into our homes and begin to capture and record all aspects of daily living, we begin to approach a world in which all, even bystanders, are **being constantly observed by various cameras wherever they go**' (Wagner, 2018).*
- *Law enforcement agencies in India already use facial recognition to predict and recognise criminal activity (Marda, 2018; Sathe, 2018)*

Concern: Freedom of speech

Freedom of speech and expression is a **fundamental right** in democratic societies that can be profoundly affected by AI.

This is a dangerous trend, given the **limited competence of ML to understand tone and context**.

AI has been widely touted by technology companies as a solution to problems such as **hate speech, violent extremism and digital misinformation** (Li and Williams, 2018).

In India, **sentiment analysis** tools are increasingly deployed to gauge the tone and nature of speech online (Marda, 2018).

The Indian Government has also expressed interest in using AI to **identify fake news and boost India's image on social media** (Seth 2017).

Harms, concerns and risks

- **Short-term**: What is the impact of [misinformation, manipulation, bias, **fairness, privacy, transparency**] of my work?
- **Mid-term**: How the use [**artistic, military, medical care, justice, education**] of these works will change the way we are organized as a society? What are the big job changes?
- **Long-term**: What are the ethical long-term **goals** of these technologies [human rights, inequality, trust]?

Concern: Impact on society: The labour market



47 % of the jobs to automate over the next 20 years:

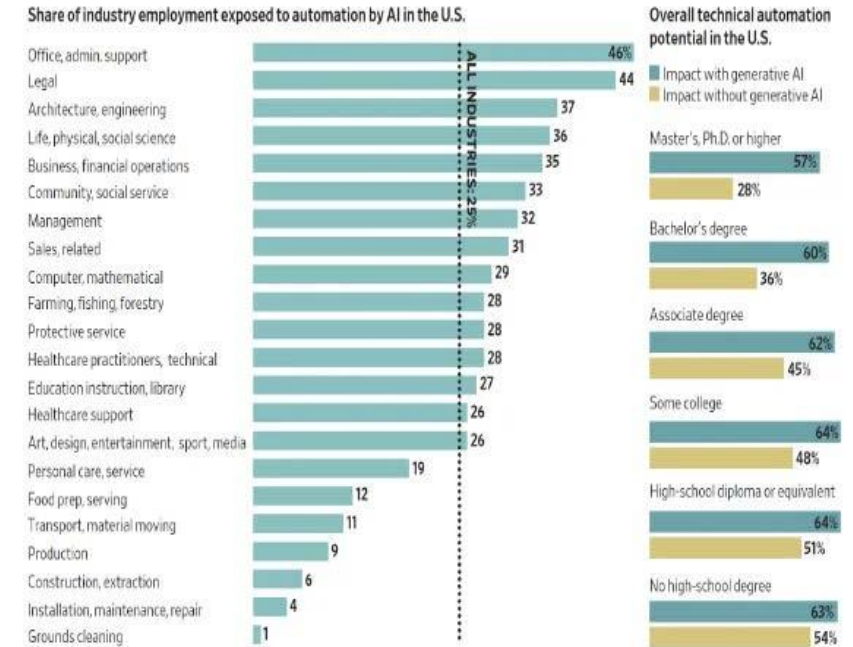
telemarketers,
title examiners,
hand sewers,
mathematical technicians,
insurance underwriters,
watch repairers,
cargo agents,
tax preparers,
photographic process workers,
new accounts clerks,
library technicians, and
data-entry specialists.



Less affected (1%):

recreational therapists,
mechanic supervisors,
emergency management directors,
mental health social workers,
audiologists,
occupational therapists,
health care social workers,
oral surgeons,
firefighter supervisors and
dieticians

Frey and Osborne (2013)



Knowledge workers and white-collar professionals are most impacted

(Source: Bob Fernandez, Wall Street Journal)



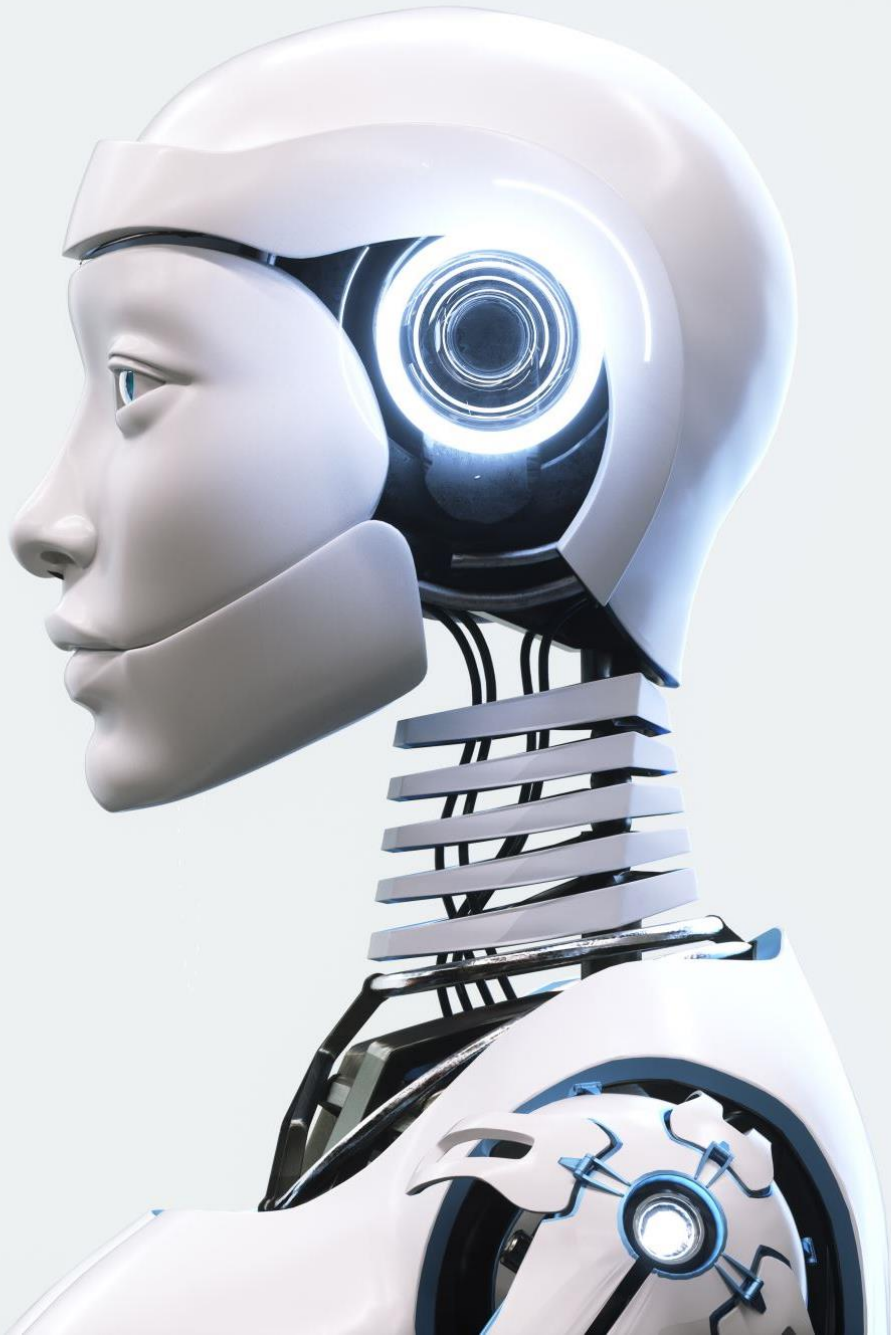
Concern: Inequality

New jobs include:

- **scanning and identifying offensive content** for deletion,
- **manually tagging objects** in images in order to create training data sets for machine learning systems
- **Interpreting queries** that an AI chatbot cannot understand.
- Many of them are **freelancers** (gig economy, ghost jobs).
- Growing 'data-labelling' industries in both China and Kenya
- Workers required to **watch and vet offensive content** for media platforms such as Facebook and YouTube

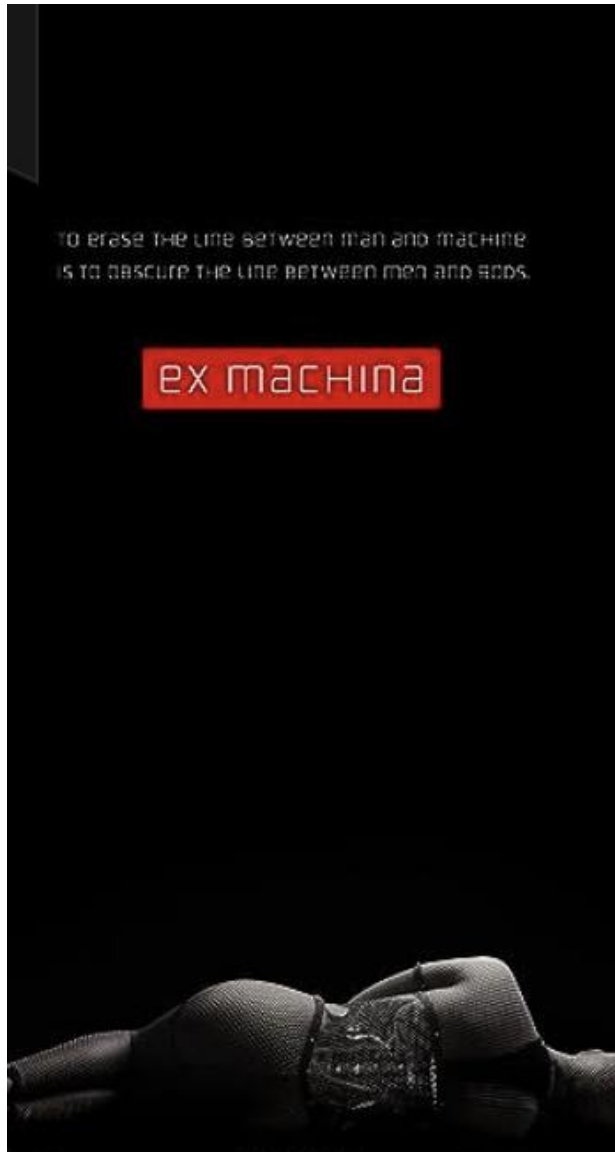
Harms, concerns and risks

- **Short-term:** What is the impact of [misinformation, manipulation, bias, **fairness, privacy, transparency**] of my work?
- **Mid-term:** How the use [**artistic, military, medical care, justice, education**] of these works will change the way we are organized as a society? What are the big job changes?
- **Long-term:** What are the ethical long-term **goals** of these technologies [human rights, inequality, trust]?



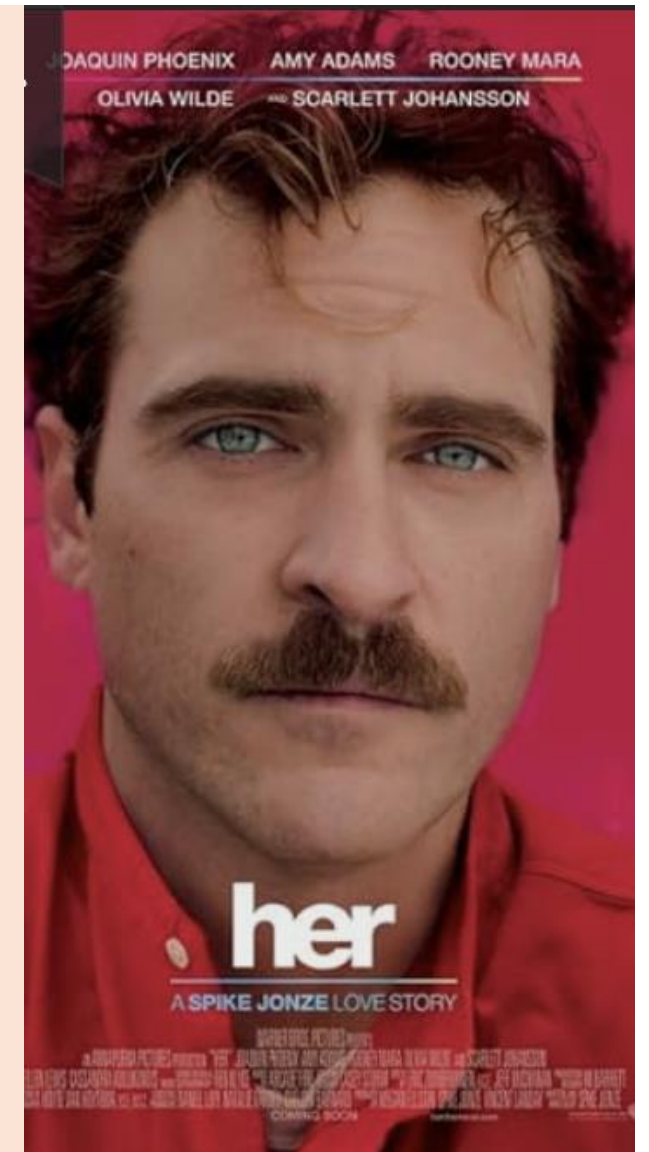
AI, Human interaction and Ethics

- The **'environment'** of a software AI is its **context**
 - clinical - medical diagnosis AI,
 - public space – face recognition in airports, or
 - virtual - face recognition in social media.
- AIs almost always **interact with humans**
 - via question and answer interfaces
 - via text for chatbots, or
 - via speech for digital assistants on mobile phones (i.e. Siri) or in the home (i.e. Alexa).
- The **interaction with humans** gives rise to almost all of the relevant **ethical issues**.



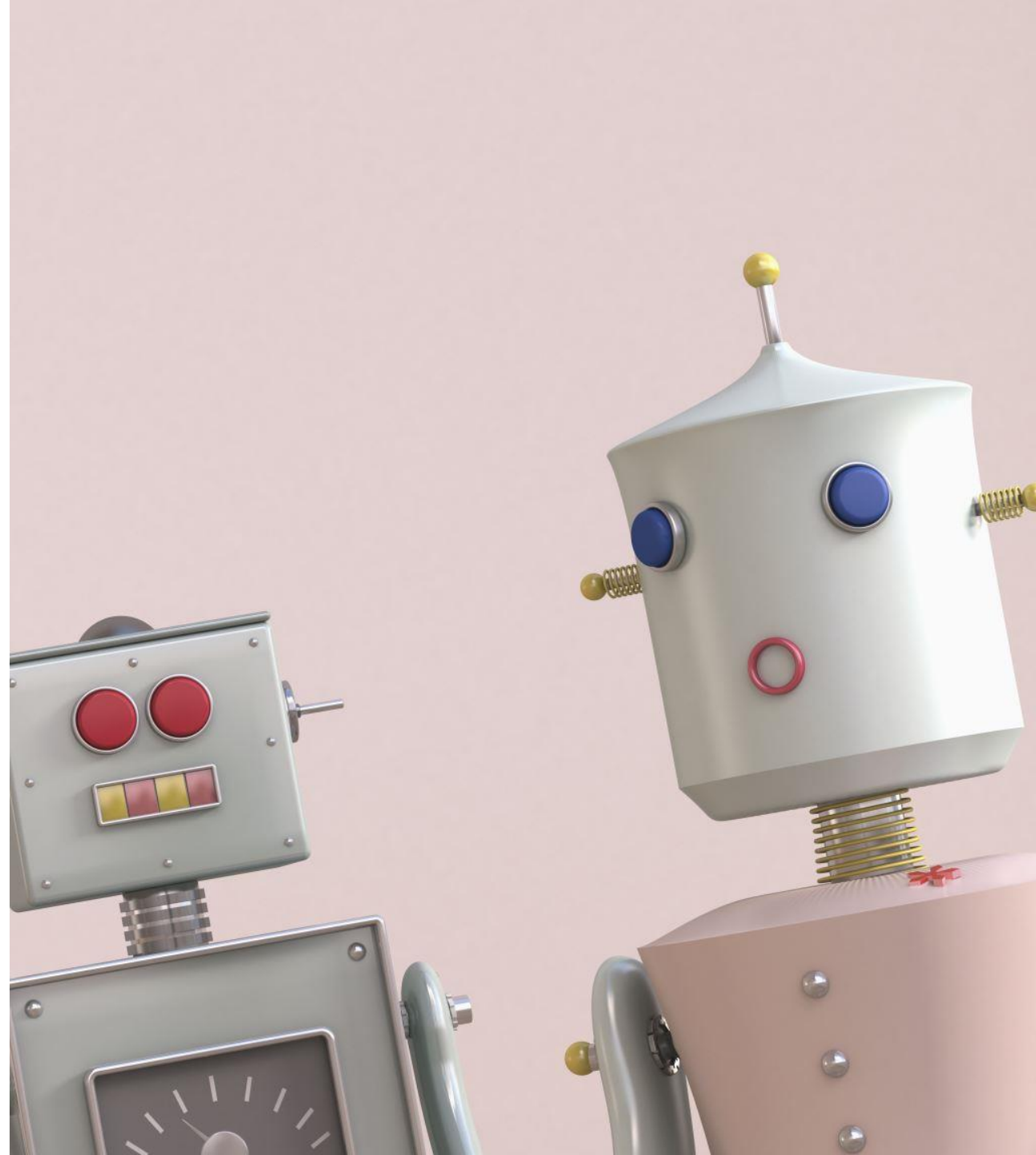
Concern: Impact on human relationship

- The biggest risk [of AI] that anyone faces is the **loss of ability to think for yourself.**
 - *"If we've lost the ability to be introspective, we've lost human agency and we're spinning around in circles." (John Havens)*
 - One danger is that of **deception and manipulation.**
 - *"Social robots that are loved and trusted could be misused to manipulate people" (Scheutz 2012)*
- It is also possible that people could become **psychologically robot-dependent.**



Concern: Human-human relationships

- Could robots **change the beliefs, attitudes, and/or values** we have about human-human relationships?
- Could 'intimate robots' lead to an **increase in violent behaviour**?
- Should we worry about the **effect of children** being rude to digital assistants such as Alexa or Siri?
 - Does this affect **how they view or treat others**?
- Research shows that robots have the capacity to **change how cooperative** we are.



Concern: Privacy, human rights and dignity

- *'Humans will not have agency and control [over their data] in any way if they are not given the tools to make it happen'. (John Havens)*
 - Personal assistants like Siri and Alexa

A survey of IPA customers showed that people's biggest privacy concern was:

- their device being hacked (68.63%),
- followed by it collecting personal information on them (16%),
- listening to their conversations 24/7 (10%),
- recording private conversations (12%),
- not respecting their privacy (6%),
- storing their data (6%) and
- the 'creepy' nature of the device (4%) (Manikonda et al, 2018).

Concern: Privacy and ethical issues



What level **of control will subjects have over the data** that are collected about them?



Should individuals have **a right to use the model, or at least to know what it is used for**, given their stake in training it?



Could ML systems seeking patterns in data **violate people's privacy** (sequencing the genome of one family member revealed health information about other members of the family)?



How to **prevent the identity, or personal information, of an individual involved in training a model** from being discovered (e.g. through a cyber-attack).

Concern: Human rights



If AI can be used to determine people's political beliefs, then individuals in our society might become **susceptible to manipulation**.



Political strategists could use this information to identify which voters are likely to be persuaded to change party affiliation, or to increase or decrease their probability of turning out to vote, and then to apply resources to persuade them to do so.

Such a strategy has been alleged to have significantly affected the outcomes of recent elections in the UK and USA (Cadwalladr, 2017).

Concern: Transparency

- Find out **why an autonomous system made a particular decision**, especially if that decision caused harm
- The **black box** problema of Deep learning algorithms
 - The AI Now Institute at New York University urged public agencies responsible for **criminal justice, healthcare, welfare and education** to **ban black box AIs** because their decisions cannot be explained.
- How to **verify the system** that fulfils specified design requirements
 - Current verification approaches typically assume that the system being verified **will never change its behaviour**,
 - But ML systems change their behaviour, so any verification is likely to be **rendered invalid** after training/adapting (Winfield and Jirotko, 2018).

Concern: Impact on Trust

- Trust in AI can only be attained by **fairness, transparency, accountability and regulation**.
- How **much control** we want to exert over AI?
- Shall we **maintain a human-in the loop**, or give systems more **autonomy**?
- People favour the idea **of prioritising the use of robots in areas that pose risk or difficulty** to humans — space exploration, manufacturing, military, security, and search and rescue
- Feel **uncomfortable with areas involving vulnerable or dependent areas** of society.
 - **to care for children, the elderly, and the disabled**; for education; and for healthcare, despite many holding positive views of robots in general.
 - **idea of having their dog walked by a robot, having a medical operation performed by a robot, or having their children or elderly parents minded by a robot** — scenarios in which trust is key.

Why trust is important?

- AI is so pervasive that trusting it is essential for our societies to work properly.
 - Constantly supervising AI algorithm used to make a decision would require significant time and resources, to the point that using digital technologies would become unfeasible
- But how do we find the correct level of trust?
- In the **short term**, **design** could play a crucial role.
 - For example, pop-up messages alerting users to algorithmic search engine results that have taken into account the user's online profile, or messages flagging that the outcome of an algorithm may not be objective.
- In the **long term**, an **infrastructure** is needed that enforces norms such as **fairness, transparency and accountability across all sectors**.



Algorithm auditors

- Larsson et al. (2019) suggest a role for **professional algorithm auditors**, whose job would be to interrogate algorithms in order to ensure they comply with pre-set standards.
 - E.g. an autonomous vehicle algorithm auditor, who could provide simulated traffic scenarios to ensure that the vehicle did not disproportionately increase the risk to pedestrians or cyclists
- The **oversight programs'** function is to **'monitor, audit, and hold operational AI programs accountable'** (Etzioni and Etzioni 2016).
 - E.g. an algorithm that conducts real-time assessments of the amount of bias caused by a news filtering algorithm, raising an alarm if bias increases beyond a certain threshold.



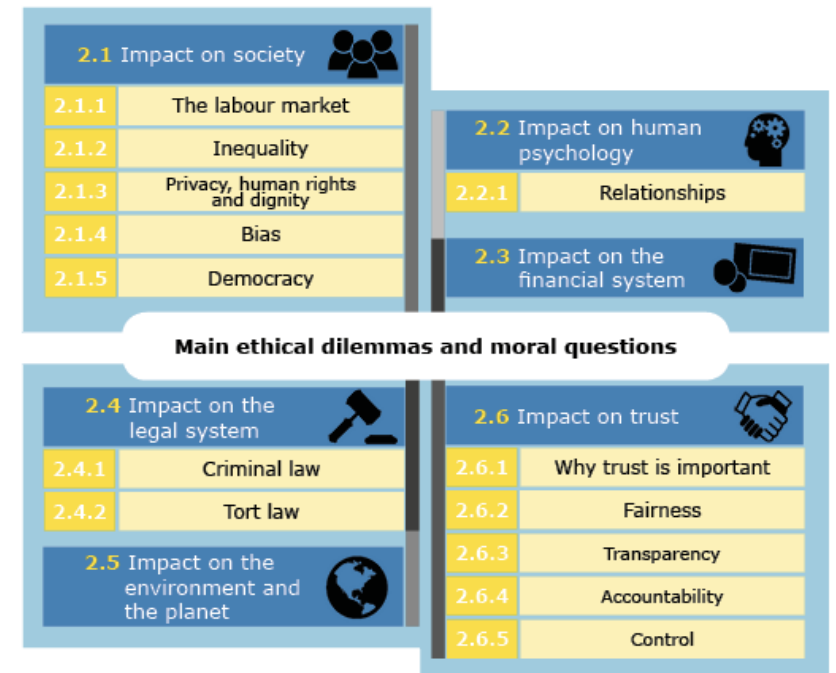
Control



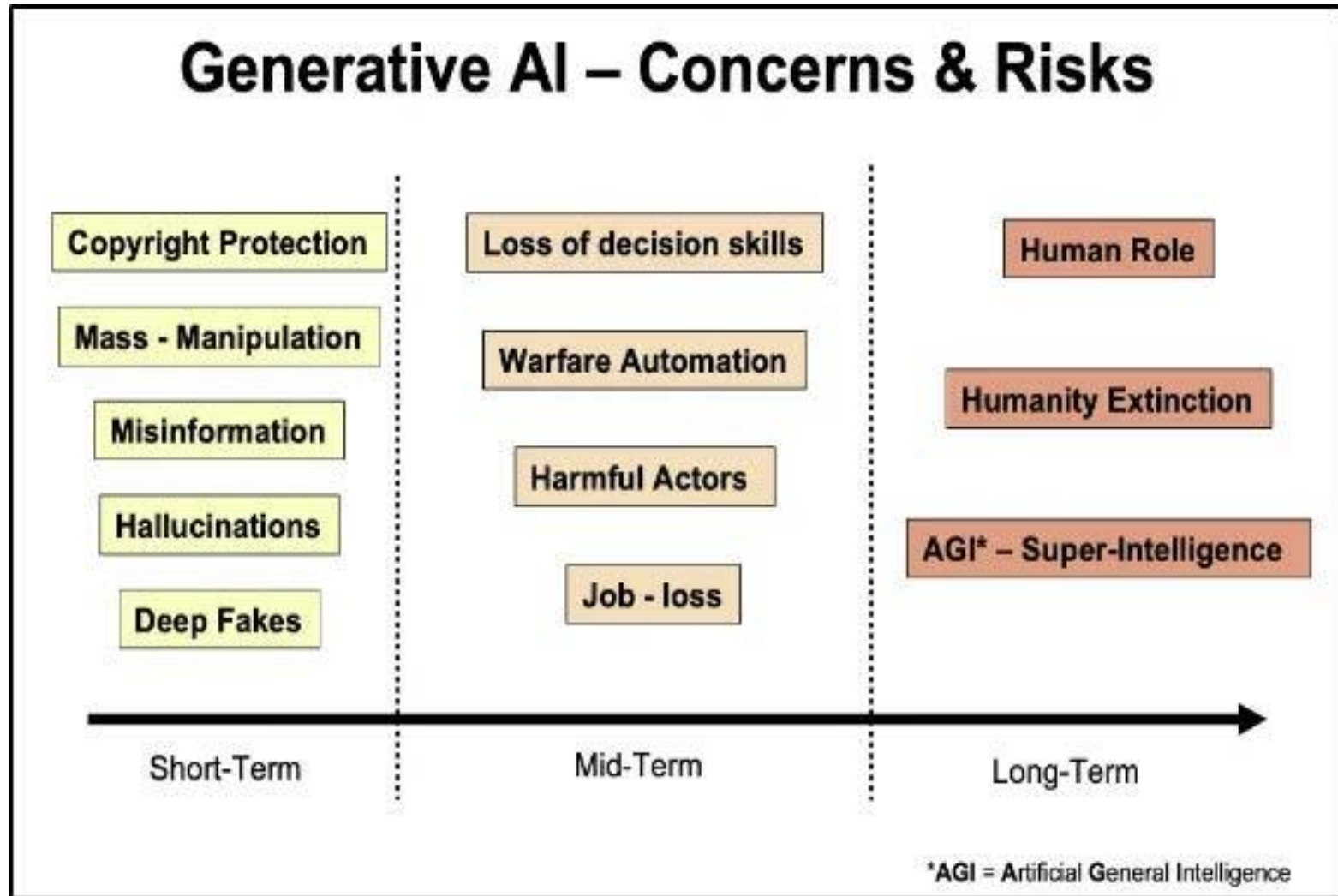
- Much of this relates to fears around the idea of 'Superintelligence' –
 - as AI surpasses human abilities, it may come to take control over our resources and outcompete our species, leading to human extinction (!?)
 - Even if an AI agent was carefully designed to have goals aligned with human needs, it might develop for itself unanticipated subgoals that are not.
- Human-in-the-loop (e.g. marking an email as spam)
 - The human can identify misbehaviour by an autonomous system, and take corrective action
 - Keeping humans in the loop would also provide accountability
 - Maybe not to be enough – society-in-the-loop.

Conclusions on main ethical and moral issues

- AI is creating one of the **most significant transformations** in human society.
 - For the first time in history, **AI is beating human expertise**.
 - It will **transform almost all jobs, professions, and industries**.
- AI is a very powerful tool with the potential **to enhance the way we live, work, make decisions, and develop solutions**.
 - Every person, business, institution, and country **can benefit** enormously through AI.
- But as with any powerful tool, it comes with significant **risks and potential for abuse** by bad actors.
- **AI requires regulations** to prevent or at least mitigate the risks,
 - We need to find new ways to enforce these regulations.



AI Concerns and Risks



Why trust is important?

*“Imagine a society in which there is **no trust in doctors, teachers, or drivers**. Without trust we would have to spend a **significant portion of our lives devoting time and resources** to making sure other people, or things were doing their jobs properly (Taddeo, 2017).”*

'We trust machine learning algorithms to indicate the best decision to make when hiring a future colleague or
- when granting parole during a criminal trial,
- to diagnose diseases and identify a possible cure,
- to drive or fly us around the globe.

This trust is widespread and is resilient. It is only reassessed (rarely broken) in the event of serious negative consequences.' (Taddeo, 2017)

Thank you!

petia.ivanova@ub.edu

