

ACCA7026

Ethics and Risks of AI

Dr. Seena Joseph

Artificial Intelligence

- AI systems have the potential to transform the society
 - It is a promising strategy for enhancing societal and individual well-being
 - Bringing progress and innovation
- Artificial intelligence (AI) is progressively influencing how we live our daily lives.
- Our lives, education, and businesses stand to benefit greatly from AI, but with tremendous power comes huge responsibility.
- AI has huge potential to bring accuracy, efficiency, cost savings and speed to a whole range of human activities.
- While offering great opportunities, AI systems also give rise to certain risks that must be handled appropriately and proportionately.
- Embedding trustworthy AI is important.
 - This means attempting to maximise AI systems' advantages while both preventing and minimising their risks.

Artificial Intelligence Applications

Health care



<https://blog.labtag.com/the-top-applications-for-ai-in-medical-diagnostics/>

Education



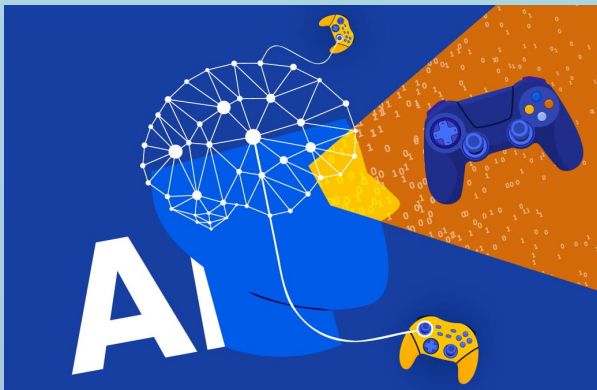
<https://yipinstitute.org/article/artificial-intelligence-in-education-helpful-or-hindrance>

Self-driving cars



<https://www.market-prospects.com/articles/artificial-intelligence-in-self-driving-cars>

Gaming



<https://pixelplex.io/blog/how-ai-enhances-game-development/>

Finance



<https://itsecuritywire.com/news/financial-services-are-banking-on-ai-implementation/>

Impact of AI in our lives and society

[AI Gone Wrong: From Deloitte to Air Canada, major global incidents exposing the dark side of technology - CNBC TV18](#)

[TOP AI incidents of 2024. 2024 has been nothing short of a... |
by Law and Ethics in Tech | Medium](#)

Impact of AI in our lives and society

Uber Self Driving Car Caused a Fatality



<https://onlinecourseing.com/ai-fails/>

Google says sorry for racist auto-tag in photo app

- Google Photos labelled a picture of two black people as 'gorillas'
- Google Maps and Flickr have also suffered from race-related problems



[Google says sorry for racist auto-tag in photo app](#) | [Google](#) | [The Guardian](#)

A lawyer used ChatGPT to prepare a court filing. It went horribly awry.

**MONEY
WATCH**

BY MEGAN CERULLO

UPDATED ON: MAY 29, 2023 / 5:40 PM / MONEYWATCH

<https://www.cbsnews.com/news/lawyer-chatgpt-court-filing-avianca/>

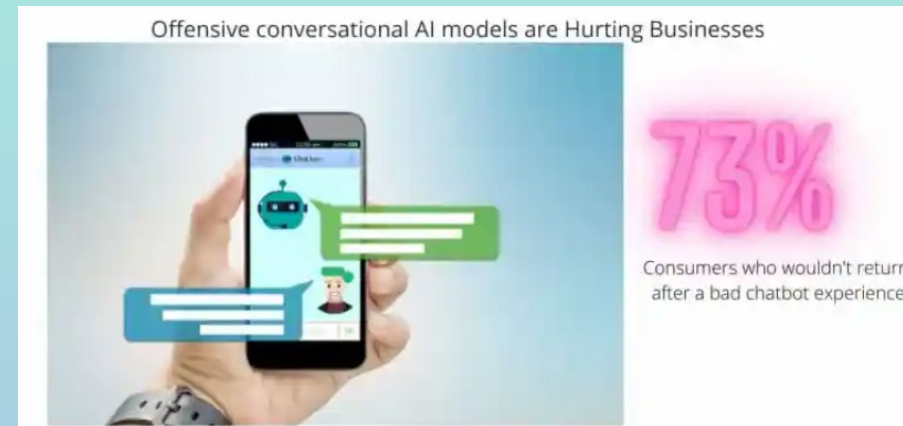
Impact of AI in our lives and society

Amazon Recruiting Software Developed Bias



<https://onlinecourseing.com/ai-fails/>

Microsoft Tay Chatbot Became Offensive



Impact of AI in our lives and society

- Automated classification systems can deliver prejudiced results and therefore raise questions about privacy and bias.
- Autonomy of self-driving vehicles raises concerns about safety and responsibility.
- Autonomous vehicles can replace a large number of jobs, and transform the transportation and associated industries.
- AI-recruiters are known to exhibit human biases because the training data inherits the same biases we have as humans.

AI Ethics

Normative Ethics	Reference
Ethics is the capacity to think critically about moral values and direct our actions in terms of such values.	Churchill, 1999
Ethics is a set of concepts and principles that guide us in determining what behavior helps or harms sentient creatures.	Paul & Elder, 2006
Ethics is the norm for conduct that distinguishes between acceptable and unacceptable behavior. Ethics is the discipline that studies standards of conduct, such as philosophy, theology, law, psychology, or sociology. Ethics is a method, procedure, or perspective for deciding how to act and for analyzing complex problems and issues.	Resnik, 2011
Applied Ethics	
Computer ethics is the analysis of the nature and social impact of computer technology and the corresponding formulation and justification of policies for the ethical use of such technology.	Moor, 1985, p. 266
Machine ethics is concerned with giving machines ethical principles or a procedure for discovering a way to resolve the ethical dilemmas they might encounter, enabling them to function in an ethically responsible manner through their own ethical decision making.	Anderson and Anderson, 2011, p. 1

Table extracted from: Siau et.al., [1]

AI Ethics

- Ethics – Principle or guidelines that help determine what is good or right.
- Discipline dealing with right versus wrong and the moral duties and obligations of entities.
- AI Ethics – subfield of philosophy that addresses the moral implications and responsibilities associated with the development and use of AI systems.
- Ensuring that AI technologies are developed and used in ways that respect human rights and promote fairness, accountability and transparency.
- AI ethics is a critical area of study that seeks to ensure that artificial intelligence is developed and used in ways that are **ethical**, **lawful**, and **robust**, ultimately contributing to a more just and equitable society

Trustworthy Artificial Intelligence

- Trustworthy AI refers to AI systems that are designed to be **lawful, ethical, and robust** throughout their lifecycle. It emphasizes the importance of meeting specific requirements to gain user trust and ensure societal benefits.
- Three pillars of trustworthy AI
 - **Lawful** – AI systems should comply with applicable laws and regulations. Should not break any law while they operate.
 - **Example** : For example, an AI system used in hospitals must comply with health regulations to protect patient information and ensure accurate diagnoses. Imagine a robot that helps doctors diagnose diseases. This robot must follow both general data protection laws and specific medical regulations to ensure it does not misuse patient data or make incorrect diagnoses.
 - **Ethical** – AI systems should act according to ethical principles and values.
 - **Example** : Use of large language models. If it is used for misleading information, it will harm the society.
 - **Deepfake** ?

Trustworthy Artificial Intelligence

- **Robust** - AI systems should be strong and reliable in both technical (performance) and social (social) perspective.
 - **Example** : Technical robustness - if an AI system is used in a self-driving car, it must be able to navigate safely without crashing.

Four Ethical principles :

1. Respect of human autonomy
2. Prevention of harm
3. Fairness
4. Explainability

Ethical principles

Respect of human autonomy

Ensuring that people have control over their own lives and decisions. AI should be a tool that supports and empowers humans, rather than one that manipulates or controls them. By focusing on these principles, we can create a future where AI enhances human capabilities and respects individual freedoms.

- Ensure the respect for the freedom and autonomy of human beings
- AI systems should be designed to augment, complement and empower human cognitive, social and cultural skills.
- Follow human centric design principles and leave meaningful opportunity for human choice.
- AI systems fundamentally alter the workplace.
- It should support humans in the working environment and aim for the creation of meaningful work

Ethical principles

The principle of prevention of harm

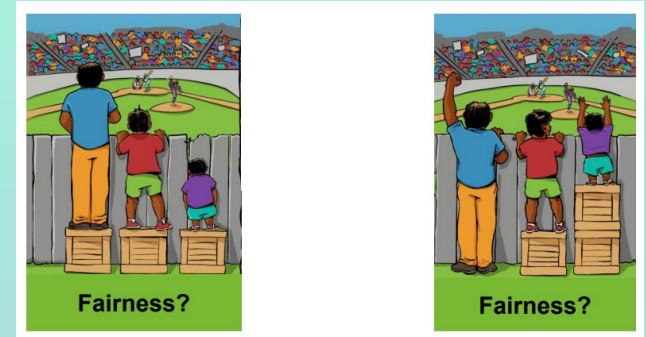
Should not adversely affect human. System should be designed to support people's mental and physical health.

- Artificial intelligence (AI) systems shouldn't damage, aggravate, or negatively impact people in any way.
- Protection of human dignity, mental and physical integrity
- AI systems and the settings in which they function need to be safe and secure.
- AI systems must be technically robust

Ethical principles

The principle of fairness

- Treating everyone equally and justly. Three important rights : Non-discrimination, solidarity and justice
- Fairness is required in the design, implementation, and use of AI systems.
 - Substantive and procedural dimensions
 - Substantive dimension implies that both individual and groups are free from unfair bias, discrimination and stigmatization.
 - Procedural dimensions entails the ability to challenge judgements made by AI systems and by the people who operate them and seek meaningful remedy.



<https://www.altexsoft.com/blog/responsible-ai/>

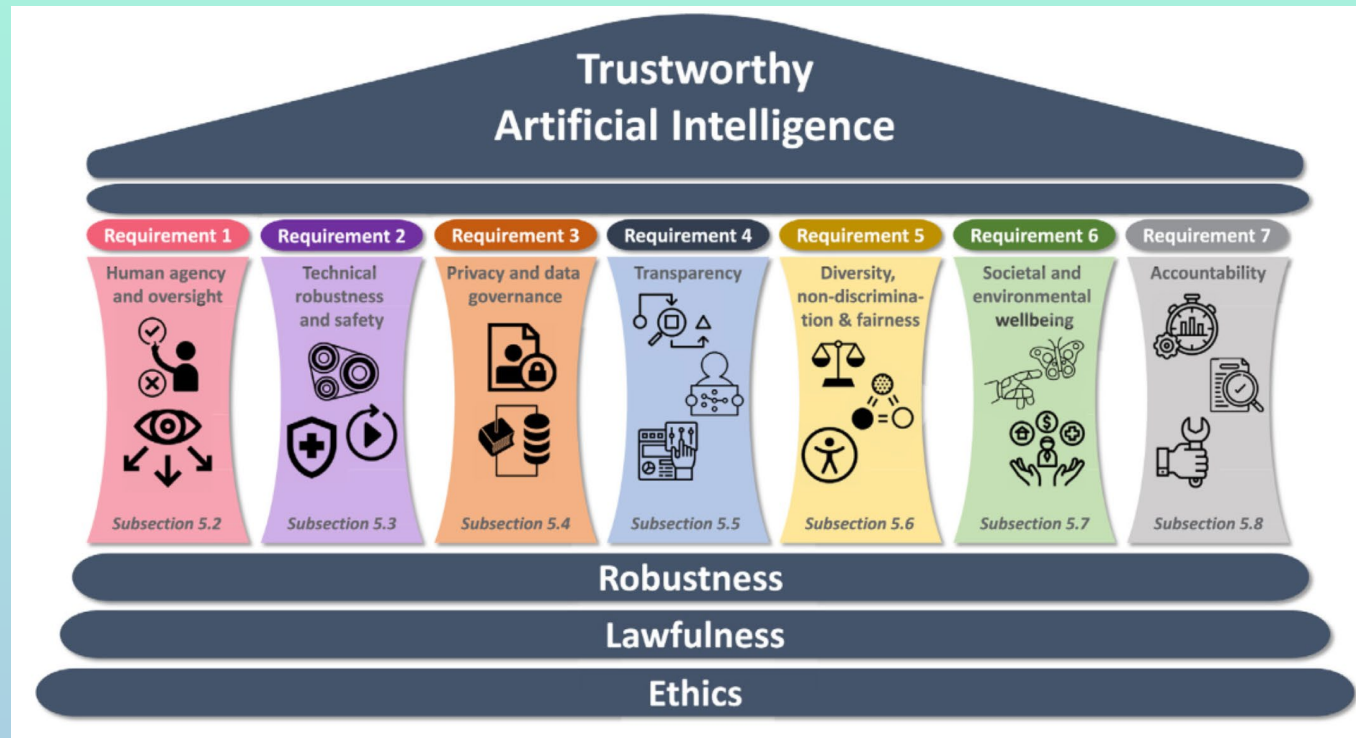
Ethical principles

The principle of explainability/explicability

- The processes need to be transparent, the capabilities and function of AI systems should be communicated
- Decision should be explainable to those directly and indirectly affected.
- It is not always possible to provide a reason for a AI system's output or choice, along with a breakdown of the input factors/parameters that affected it. E.g. *“black box” algorithm [mysterious to even developers]. In these cases, other measures such as traceability, auditability and transparent communication on systems capability is required.*
- *The level of explainability necessary depends on the situation and how serious the implications might be in the event that the output is incorrect or erroneous.*

Trustworthy AI

Trustworthy AI



Source [2]

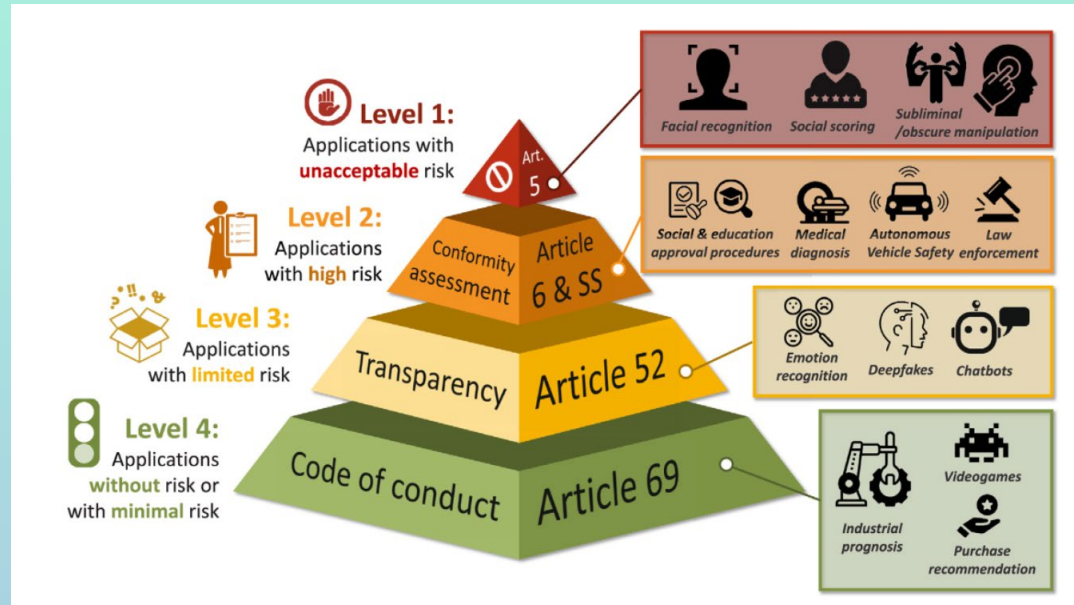
Trustworthy AI

Trustworthy AI – Requirements

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental wellbeing
7. Accountability

Artificial Intelligence Regulation

A risk-based approach - the most advanced regulation to date is the European Commission's AI Act proposal



Source [2]

- [AI Act | Shaping Europe's digital future \(europa.eu\)](#)

Artificial Intelligence Regulation

Minimal or no-risk

AI systems that pose a low likelihood of causing harm to individuals or society.

Example : ??

Limited risk

Limited risk refers to AI systems that have a moderate potential to cause harm but are not classified as high-risk. These systems may have some implications for safety, privacy, or fundamental rights, but the risks they pose are manageable and do not require the same level of regulatory scrutiny as high-risk systems .

Example : ??

High risk

Systems that can have a significant impact on the life chances of a user.

Example : ??

Unacceptable risk

Unacceptable risk refers to AI systems that pose a clear and significant threat to the safety, livelihoods, and rights of individuals. These systems are deemed too dangerous to be allowed in the market due to their potential for causing harm or violating fundamental rights.

Example : ??

Responsible AI

Responsible AI: develop, assess, deploy AI systems in a safe, trustworthy and ethical way.

Fairness, Transparency, Accountability, Privacy and Security

Fairness

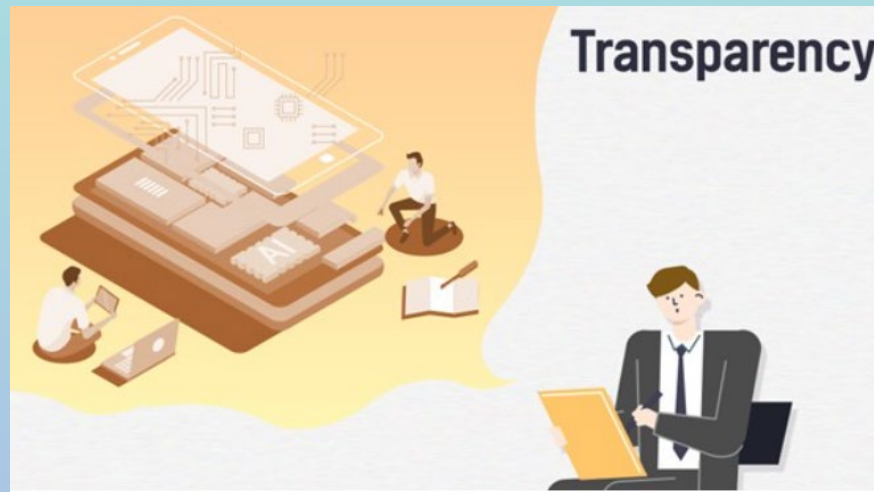
Ensuring that AI algorithms do not discriminate against any individual or group. AI systems should be developed to reduce bias and discrimination based on race, gender, age, religion, or other protected characteristics.

Responsible AI

Transparency

Making AI systems understandable and providing clear explanations for their decisions.

- Traceability (the data, preprocessing steps, and model is accessible)
- Explainability (the reasoning behind decision-making/prediction is clear)
- Open Communication (regarding the limitation of the AI system)



Responsible AI

Accountability

Holding developers and users responsible for the AI's actions and consequences.

Privacy and Security

Safeguarding personal data and ensuring AI systems are protected against potential threats.

Responsible use of AI

- Responsible approach to AI is needed to ensure the safe, beneficial and fair use of AI technologies.
- The responsible approach to AI not only ensures that the systems are developed in a good way, but also that they are developed for a good cause.

Responsible use of AI

- Example: Responsible use of chatGPT (chat Generative Pre-trained Transformer)



Fairness

- Bias in training data
 - Developers: monitor and evaluate the data
 - Users: Being critical about the response and cross-referencing information

Transparency

- “Blackbox “
 - Developers: Explainable AI Tool, communicate the limitation of the model
 - Users: ensure accurate and transparent information by verifying answers and seek additional contexts

Accountability

- Developers: misuse or negative outcome associated with it
- Users: understand the limitation and do not rely on the model for the critical task.

Privacy and security

- Developers: Robust data protection mechanism, safeguard user data, secure to measures to prevent a potential attack
- Users: Cautious about sharing private and confidential information,

Responsible use of AI

- Case Study - Read about
 - Recent Autonomous Vehicle (AV) Accident
 - Recent incidents of AI Hallucinations (e.g. People or professionals using AI-generated false information)
- Tasks
 - Summarize what happened.
 - Identify the ethical issues and risks involved.
 - Suggest how such incidents could be prevented.

Further reading

- [Why the biggest challenge facing AI is an ethical one - BBC Future](#)
- N. Díaz-Rodríguez, et al., "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Information Fusion*, vol. 99, 2023. DOI: 10.1016/j.inffus.2023.101896
- It's time to talk about the known risks of AI. *Nature* 618618(7967), pp. 885-886. 2023. . DOI: 10.1038/d41586-023-02094-7.

References

- [1] K. Siau and W. Wang. K. Siau and W. Wang. Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. Artificial intelligence (AI) ethics: Ethics of AI and ethical AI. *J. Database Manage. J. Database Manage.* 3131(2), pp. 74-87. pp. 74-87. 2020. . DOI: 10.4018/JDM.2020040105.
- [2] N. Díaz-Rodríguez, et al., "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Information Fusion*, vol. 99, 2023. DOI: 10.1016/j.inffus.2023.101896.