

Algorithmic foundations and ethics in AI: from theory to practice course

Toolkit for synchronous sessions

CU1 | AI ethics - a practical approach
Support PowerPoint slides

INDEX

- INTRODUCTION - 3
- ETHICAL PRINCIPLES OF AI - 8
- ETHICAL FRAMEWORKS, GUIDELINES AND TOOLKITS OF AI - 23
- EU AI ACT – 33
- AI DEVELOPMENT LIFECYCLE – 40
- STAKEHOLDER ENGAGEMENT IN AI SOLUTION DEVELOPMENT – 45
- ETHICAL PRINCIPLES AND STAKEHOLDER ENGAGEMENT IN AI DEVELOPMENT LIFECYCLE - 53

INTRODUCTION

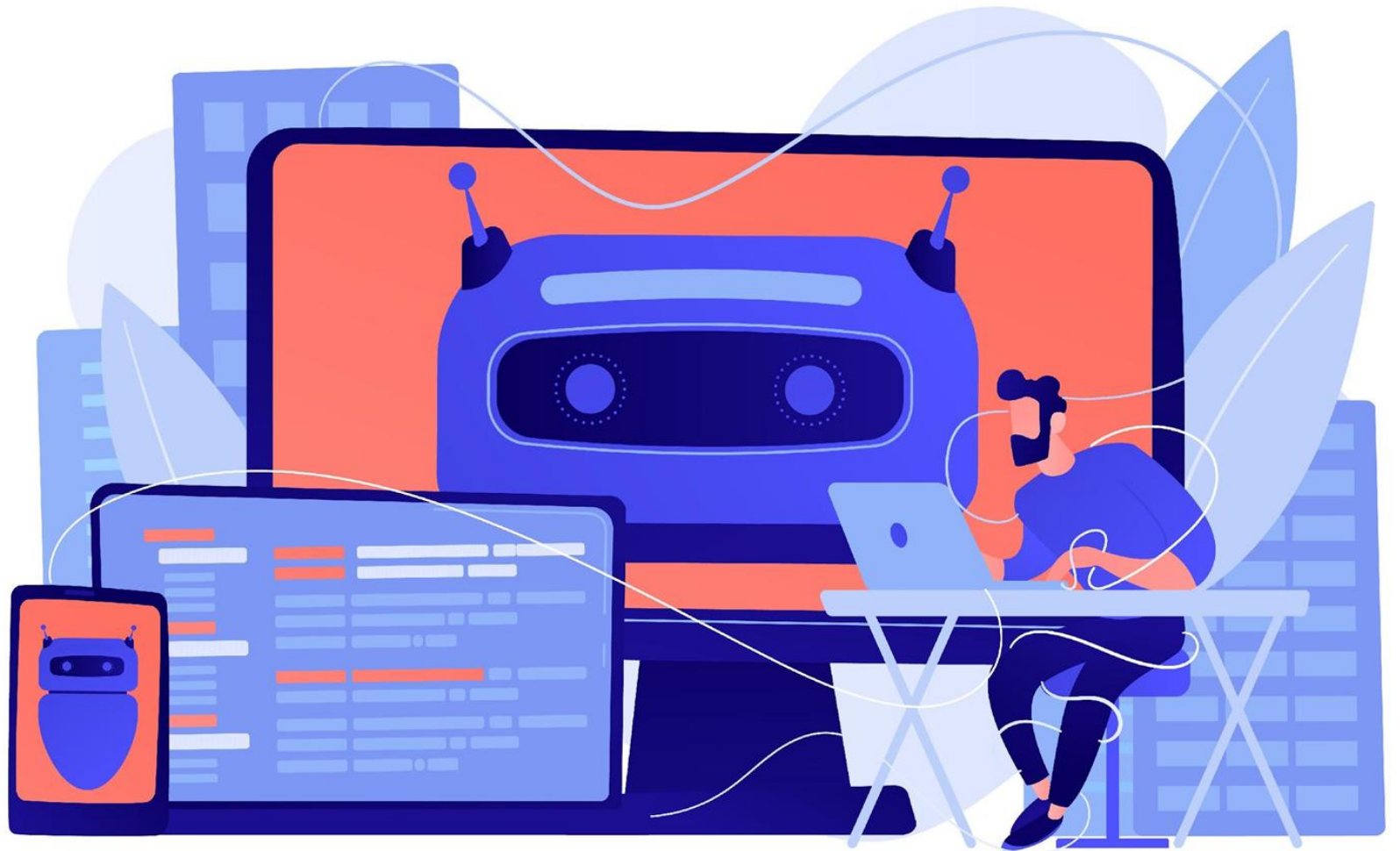


IMAGE SOURCE | Freepik

IN THIS COMPETENCE UNIT YOU WILL FIND THE FOLLOWING SUBJECTS:

- AI ethics and its importance in AI solution development
- Ethical principles of AI
- Ethical frameworks, guidelines and toolkits
- High level expert group; Trustworthy AI – Framework
- EU AI act
- AI development lifecycle
- Stakeholder engagement in AI development lifecycle
- AI development lifecycle with related ethical principles and stakeholders

AT THE END OF THE COMPETENCE UNIT, YOU SHOULD BE ABLE TO:

- Be aware of the ethical principles in the AI development and deployment
- Recognize various ethical frameworks and guidelines and their application in real world scenarios involving AI systems
- Understand the importance of ethical principles and stakeholder engagement in the ethical AI development process

Ethical concerns – What happens if...

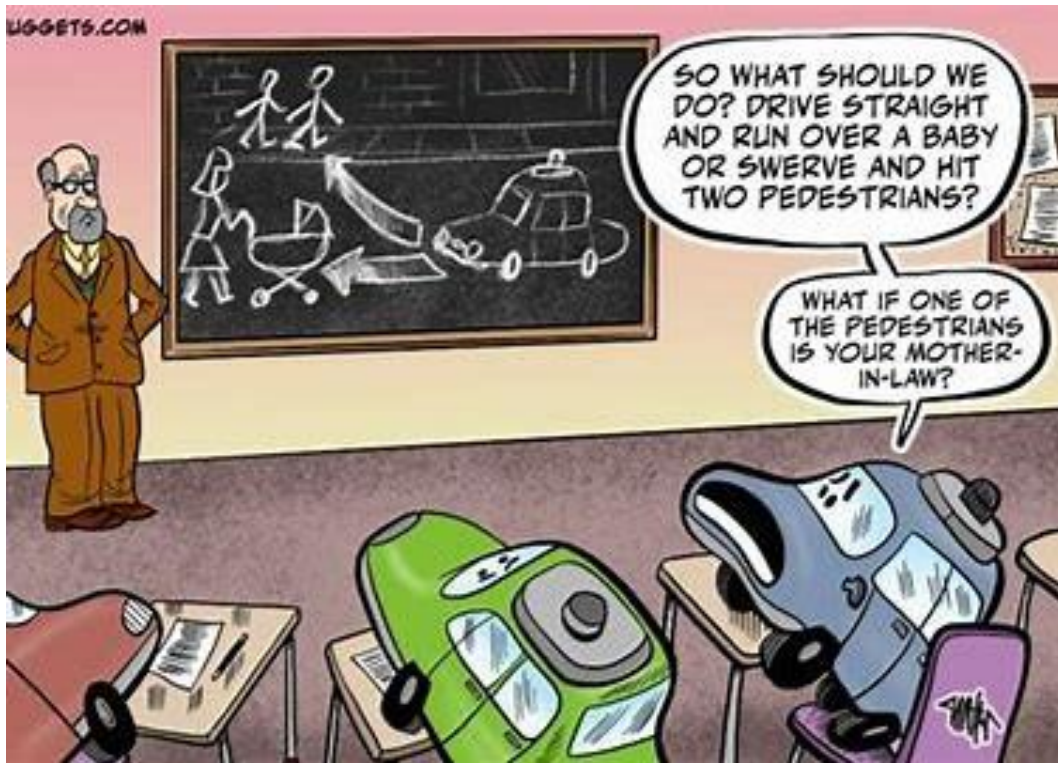


IMAGE SOURCE | [Cartoon: Teaching Ethics to AI - KDnuggets](#)

- AI would have the power to decide whether you get a loan?
- AI would judge who will be sentenced for prison?
- An autonomous car would have to decide whether to save the life of the pedestrian or the lives of those inside the car?
- What if an AI-based recruitment system would decide whether to hire a female or male candidate as a flight captain?

Artificial intelligence is revolutionizing life



IMAGE SOURCE | OECD; AI: Risk and Opportunity

Video from: <https://youtu.be/-CXkHs3cxa4>

- AI impacts all facets of life, influencing work, leisure, and providing solutions to global issues like climate change and healthcare access, while also posing significant challenges for governments and individuals.
- The integration of AI into economies and societies raises questions about the appropriate policy and institutional frameworks needed to guide its development and application for societal benefit.
- Broad ethical considerations have led to the creation and publication of numerous approaches to ensure the ethical application of AI, which are covered in this course.

ETHICAL PRINCIPLES OF AI



Ethical principles of AI

What are ethical principles of AI?

- **Ethical principles are defined as general guidelines that address ethical concerns of AI systems.** These principles are often abstract and do not provide specific instructions on how to apply them in a particular context.
- Ethical principles describe what is expected in terms of right and wrong and other ethical standards. Ethical principles of AI refer to normative constraints on the **“do’s” and “don’ts”** of algorithmic use in society.

(Prem 2023), (Zhou et al., 2020)

Ethical principles of AI

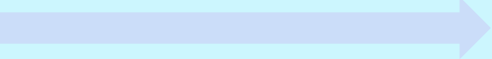
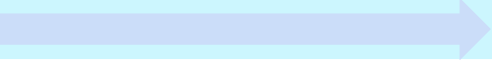
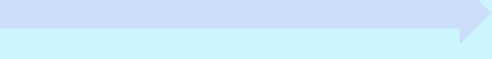
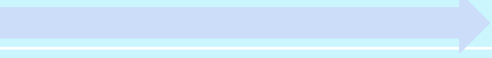

- Multitude of recommendations has been published to guide the ethical use of artificial intelligence.
- Contributions come from various sectors eg:
 - OECD's Recommendation, outline AI ethics principles (Published May 2019)
 - Beijing AI Principles for R&D (Published May 2019)
 - Industry (e.g., Google, IBM, Microsoft, Intel)
 - Government (e.g., Montreal Declaration, European Commission's Expert Group)
 - Academia (e.g., Future of Life Institute, IEEE, AI4People)

(Prem 2023), (Morley et al 2020)

Ethical principles of AI

Examples of ethical principles

The graph is showing how different frameworks combine ethical principles and how they overlap. Horizontal arrows link the topics to each other.

AI4People (published November 2018) (Floridi et al. 2018)	Five principles key to any ethical framework for AI (L Floridi and Clement-Jones 2019)	Ethics Guidelines for Trustworthy AI (Published April 2019) (European Commission 2019)	Recommendation of the Council of Artificial Intelligence (Published May 2019) (OECD 2019b)
Beneficence 	AI must be beneficial to humanity	Respect for human autonomy	Inclusive growth, sustainable development and well-being
Non-Maleficence 	AI must not infringe on privacy or undermine security	Prevention of harm	Robustness, security and safety
Autonomy 	AI must protect and enhance our autonomy and ability to take decisions and choose between alternatives		Human-centred values
Justice 	AI must promote prosperity	Fairness	Fairness
Explicability 	AI systems must be understandable and explainable	Explicability	Transparency and explainability Accountability

Modified from Morley et al 2020

Ethical principles of AI

Shared Principles in Ethical AI Documents



IMAGE SOURCE | Image developed by VAMK

Ethical AI principles show significant overlap, suggesting a convergence toward a unified set of ethical guidelines.

- Beneficial and respectful to people and the environment (**beneficence**).
- Robust and secure (**non-maleficence**).
- Respectful of human values (**autonomy**).
- Fair (**justice**).
- Explainable, accountable, and understandable (**explicability**).

Morley et al (2020), Jobin et al (2019)

Ethical principles of AI



IMAGE SOURCE | Freepik

Beneficence

Beneficence in AI encompasses fostering humanity's welfare

The principle of beneficence emphasizes the creation of AI that benefits people and the planet, highlighting the importance of promoting well-being and environmental sustainability.

(Floridi et al 2018)

Ethical principles of AI



IMAGE SOURCE | Freepik

Example on principle Beneficence

Healthcare recommendations

- Consider an AI-driven healthcare system that recommends treatment plans based on patient data. If the AI prioritizes treatment options that are more profitable for the healthcare provider over those that are most beneficial for the patient's health, it will fail to uphold beneficence. In this scenario, the AI's recommendations prioritize financial gains rather than maximizing the patient's well-being.
- IBM's Watson for Oncology faced ethical challenges as it recommended treatments not always aligned with medical guidelines or patient interests, raising concerns about prioritizing profit over patient outcomes. For more information on the case, visit [this link](#).

(Floridi et al 2018)

Ethical principles of AI



IMAGE SOURCE | Freepik

Non-maleficence

Non-maleficence in AI focuses on avoiding harm.

Non-maleficence focuses on preventing harm, whether from intentional human actions or unintended machine behaviors, including unintentional influence on human behavior. It's about avoiding negative outcomes regardless of the source.

(Floridi et al 2018)

Ethical principles of AI



IMAGE SOURCE | Freepik

Example on principle Non-maleficence

Predictive policing

- Imagine an AI algorithm used in predictive policing that disproportionately targets minority communities due to biases in the data it was trained on. This AI system could inadvertently perpetuate harm and injustice by subjecting innocent individuals to increased surveillance or scrutiny based on flawed assumptions. Despite its intention to reduce crime rates, the AI's actions may lead to the unjust harassment or targeting of certain demographic groups.
- The use of predictive policing software "Palantir" by the New Orleans Police Department disproportionately targeted minority communities, which raised concerns about perpetuating systemic biases and injustice in law enforcement practices. For more information on the case, visit [this link](#).

Bias in algorithms – Artificial intelligence and discrimination (europa.eu)

Ethical principles of AI



IMAGE SOURCE | Freepik

Autonomy

Autonomy in AI involves a careful balance between human and machine decision-making, emphasizing the need to preserve human self-determination.

- Autonomy involves balancing the decision-making power between humans and AI. It upholds the intrinsic value of human choice for significant decisions, advocating for a 'decide-to-delegate' model where humans retain the ultimate authority to delegate and, if necessary, override AI decisions.

(Floridi et al 2018)

Ethical principles of AI



IMAGE SOURCE | Freepik

Example on principle Autonomy

Social Media Manipulation

- An example of AI undermining autonomy could be a social media recommendation algorithm that utilizes personalized content to manipulate users' behavior and opinions without their explicit consent. By continuously feeding users with content that aligns with their existing beliefs or biases, the AI restricts users' exposure to diverse perspectives and influences, thus limiting their autonomy to make informed decisions based on a wide range of information.
- **Real-life example:** Facebook's News Feed algorithm, designed to personalize content based on users' interactions, has been criticized for fostering filter bubbles and echo chambers, potentially amplifying sensational or polarizing content and influencing user behavior without explicit consent. For more information on the topic, visit [this link](#).

Ethical principles of AI



IMAGE SOURCE | Freepik

Justice

The principle of justice in AI deals with equity in decision-making and its societal distribution, aiming to address disparities and promote fairness.

- The principle of justice is concerned with using AI to address historical injustices, ensuring equitable distribution of AI's benefits, and preventing new harms or disruptions to social solidarity.

(Floridi et al 2018)

Ethical principles of AI



IMAGE SOURCE | Freepik

Example on principle Justice

Criminal Sentencing Algorithms

- AI-driven algorithms used in criminal sentencing may unintentionally perpetuate biases against certain demographic groups. If historical data used to train these algorithms reflect biased decisions made by humans, the AI system may recommend longer sentences or harsher punishments for individuals from marginalized communities, exacerbating existing inequalities in the criminal justice system.
- The COMPAS system, utilized for risk assessment in criminal sentencing, was found to exhibit racial bias, disproportionately labeling Black defendants as higher risk and contributing to unjust sentencing outcomes. For more information on the case, visit [this link](#).

Megan T. Stevenson and Christopher Slobogin, University of Pennsylvania Law Review

Ethical principles of AI



IMAGE SOURCE | Freepik

Explicability

Explicability in AI means making sure we understand how AI makes its decisions and being able to explain those decisions clearly. It ensures that AI systems are transparent and accountable for their actions.

(Floridi et al 2018)

Ethical principles of AI



IMAGE SOURCE | Freepik

Example on principle Explicability

Credit Scoring Algorithms

- Consider an AI-powered credit scoring system that determines individuals' creditworthiness but lacks transparency in its decision-making process. If applicants are denied loans or offered unfavorable terms without understanding the factors considered by the AI algorithm, it undermines their ability to challenge or appeal the decisions. Without clear explanations of how the AI assesses credit risk, individuals may feel unfairly treated and lose trust in the system's integrity.
- The Equifax data breach in 2017 exposed millions of consumers' sensitive information, raising concerns about the lack of transparency and security in credit scoring algorithms, highlighting broader industry issues regarding consumers' limited visibility into factors impacting creditworthiness. For more information on the case, visit [this link](#).

Keith Frankish & William M. Ramsey

ETHICAL FRAMEWORKS, GUIDELINES AND TOOLKITS OF AI



Ethical frameworks, guidelines and toolkits of AI

Ethical frameworks

- **Ethical frameworks go beyond high-level statements of principles. They provide a structured approach for ethical decision-making.**
- Many frameworks for ethical AI aim to identify potential ethical challenges and propose some remedies to overcome those challenges or mitigate the associated risks.
- An ethical framework for AI serves as a foundational structure for shaping laws, rules, technical standards, and best practices across various sectors and regions.
- Typical components of ethical frameworks are:
 - Concepts – Relevant for debating the ethical aspects
 - Principle – Ethical principles (e.g. values)
 - Concern – How principles are threatened through AI systems use and development
 - Remedy – Strategies, rules and guidelines for addressing the concern

(Floridi & Cowls, 2019), (Prem, 2023: 701) (Ayling & Chapman, 2022)

Ethical frameworks, guidelines and toolkits of AI

Ethical toolkits & guidelines

- **Ethical guidelines offer practical guidance and specific rules for ethical behavior in various contexts.**
- Ethical **toolkits and guidelines** shape AI-based innovation and support the **practical application of ethical principles of AI.**
- Toolkits and guidelines explain how to apply ethical principles into the design, implementation, and deployment of AI.
- Frameworks, guidelines, and toolkits frequently overlap in their use.
- In the following pages, two examples of frameworks are presented more in details; HLEG Trustworthy AI Framework and OECD AI Classification Framework

(Zhou et al., 2020)

Ethical frameworks, guidelines and toolkits of AI

High-Level Expert Group on Artificial Intelligence; Trustworthy AI



- Despite the agreement that AI should be ethical, there is debate over what constitutes "ethical AI" and what ethical requirements and technical standards are needed to achieve it.
- The European Commission, presented a strategy for AI development in Europe that emphasizes ethical, secure, and advanced AI.
- The Ethics Guidelines for Trustworthy AI were presented by the AI-HLEG on April 2019 to promote Trustworthy AI.
- **Trustworthy AI** is recognized as the essential goal for fostering confidence in AI's development and application, predicated on a robust framework ensuring its trustworthiness.
- See more from <https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai>

(Leijnen et al., 2020), HLEG (2019). High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. Brussels: European Commission

Ethical frameworks, guidelines and toolkits of AI

Trustworthy AI - Framework

Trustworthy AI has three components, which should be met throughout the system's entire life cycle:

- Lawful - respecting all applicable laws and regulations
- Ethical - respecting ethical principles and values
- Robust - both from a technical perspective while taking into account its social environment

Ethical frameworks, guidelines and toolkits of AI

High-level expert group on artificial intelligence: framework for trustworthy AI

Foundations of Trustworthy AI

- Four **ethical principles** based on fundamental rights



Realization of Trustworthy AI

- Implement seven **key requirements**



Assessment of Trustworthy AI

- For operationalizing the key requirements, tailor **assessment list** to the specific AI application

HLEG (2019). High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. Brussels: European Commission.

Ethical frameworks, guidelines and toolkits of AI

Foundations of Trustworthy AI: Ethical principles

Respect for Human Autonomy

- AI should enhance self-determination, avoid manipulation, and support human cognitive, social, and cultural skills. It must allow for human oversight and meaningful choice, and support humans in the workplace.

Prevention of Harm

- AI must not cause harm, safeguard human dignity, be secure and robust, and give special consideration to vulnerable groups and the environment. It should prevent adverse effects due to power or information imbalances.

Fairness

- AI should ensure equitable benefit and cost distribution, prevent bias and discrimination, and promote equal opportunity. It requires balance in means and ends and allows for contestation and redress against AI decisions.

Explicability

- AI processes must be transparent, with clear communication of capabilities and purposes. When full explainability isn't possible, other measures like traceability and auditability should ensure rights respect.

- AI HLEG lists four ethical principles, rooted in fundamental rights, which must be respected in order to ensure that AI systems are developed, deployed and used in a trustworthy manner.
- Ethical principles are specified as ethical imperatives, such that AI practitioners should always strive to adhere to them.
- AI HLEG also acknowledge and address the potential tensions between these principles.

<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Ethical frameworks, guidelines and toolkits of AI

Realisation of Trustworthy AI: Key requirements

- The ethical principles established must be converted into definitive **requirements** for the realization of Trustworthy AI. These requirements pertain to all participants in the AI system's life cycle:
- Developers: individuals or teams conducting AI research, design, or development.
- Deployers: organizations implementing AI in their operations and offering AI-driven products or services.
- End-users: people who interact with AI systems, whether directly or indirectly.
- Broader society: those who are impacted by AI, both directly and indirectly.

Human agency and oversight

- fundamental rights, human agency and human oversight

Technical robustness and safety

- resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility

Privacy and data governance

- respect for privacy, quality and integrity of data, and access to data

Transparency

- traceability, explainability and communication

Diversity, non-discrimination and fairness

- the avoidance of unfair bias, accessibility and universal design, and stakeholder participation

Societal and environmental wellbeing

- sustainability and environmental friendliness, social impact, society and democracy

Accountability

- auditability, minimisation and reporting of negative impact, trade-offs and redress.

HLEG (2019). High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. Brussels: European Commission

Ethical frameworks, guidelines and toolkits of AI

Assessment of Trustworthy AI: ALTAI assessment list

- To aid developing Trustworthy AI, a tool has been developed that transforms these AI principles into practical measures.
- Known as the Assessment List for Trustworthy AI (ALTAI), this tool provides a comprehensive and dynamic checklist that serves as a roadmap for developers and deployers to integrate these principles into their AI applications.
- ALTAI facilitates this process by offering a set of concrete steps for self-assessment, thereby ensuring that AI technologies are developed in a manner that maximizes user benefits while minimizing exposure to unnecessary risks.
- Access to ALTAI [tool](#)



IMAGE SOURCE | ALTAI

(High-Level Expert Group on AI (AI HLEG), 2020)

Ethical frameworks, guidelines and toolkits of AI

OECD AI classification framework

- The OECD's AI classification framework, created by the OECD.AI Network of Experts, is designed to assist various stakeholders, in evaluating the potentials and risks of diverse AI systems.
- This tool facilitates the development of AI strategies and aims to maintain policy coherence internationally.
- The Framework is grounded in the OECD AI Principles, endorsing values like fairness, transparency, safety, and accountability, as well as advocating for human capacity development and international collaboration.



IMAGE SOURCE | OECD Framework

<https://www.oecd.org/digital/artificial-intelligence/>

EU AI ACT



EU AI Act

- The AI Act is an EU law on AI Artificial Intelligence Act | AI Act - the first of its kind in the world.
- It applies to the development, deployment, and use of AI in the EU or when it will affect people in the EU.
- EU AI act was taken into use gradually. See the recent updates from here: [EU AI Act: first regulation on artificial intelligence | Topics | European Parliament](#)
- The Act is the first-ever legal framework that sets out harmonised rules for the development, placing on the market, and use of artificial intelligence in the European Union.



IMAGE SOURCE | [The AI Act Explorer | EU Artificial Intelligence Act](#)

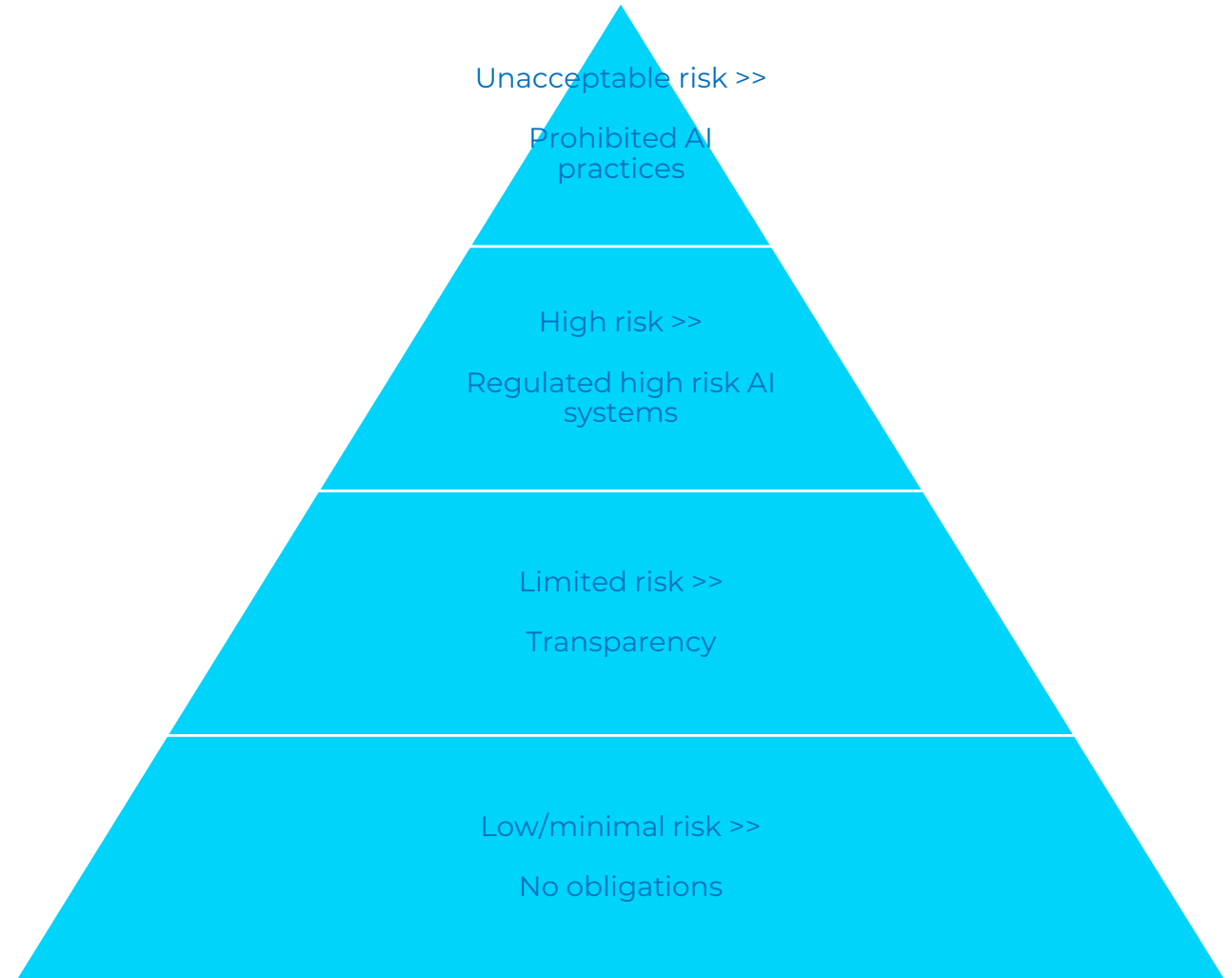
<https://artificialintelligenceact.com>
<https://www.euaiact.com>

EU AI Act

Risk-based framework

The ethical principles established must be converted into definitive **requirements** for the realization of Trustworthy AI. These requirements pertain to all participants in the AI system's life cycle:

- **Developers:** individuals or teams conducting AI research, design, or development.
- **Deployers:** organizations implementing AI in their operations and offering AI-driven products or services.
- **End-users:** people who interact with AI systems, whether directly or indirectly.
- **Broader society:** those who are impacted by AI, both directly and indirectly.



The EU AI Act 'Pyramid of Risks' (Source: European Parliament)

Risk-based framework | Minimal or no risk



AI applications that are considered to present **minimal or no risk** to citizens' rights or safety. The vast majority of AI systems fall into the category of minimal risk.

Examples of minimal risk AI applications:

- AI-enabled recommender systems
- Spam filters

Under the current regulations, these systems are not subject to any specific new obligations but must adhere to pre-existing laws.

Risk-based framework | Limited risk



Limited risk AI systems require transparency.

For these AI systems, the requirements pertain to transparency: Users should be informed that they are interacting with AI and have the necessary information to decide on continued use.

Examples of limited risk AI applications:

- AI that generates or manipulates image or audio content.
- AI that creates or alters video content, such as deepfakes.

Risk-based framework | High risk



High risk AI systems can have a significant impact on the life chances of a user.

These systems have stringent requirements to be followed before being deployed on the EU market, including risk management and data governance obligations.

High-risk AI examples include:

- Management and operation of critical infrastructure
- Education and vocational training
- Employment, worker management and access to self-employment
- Access to and enjoyment of essential private services and public services and benefits
- Law enforcement
- Migration, asylum and border control management
- Assistance in legal interpretation and application of the law.

Risk-based framework | Unacceptable risk



AI systems with **unacceptable risk** are considered to have a clear threat to the fundamental rights of people are prohibited. Unacceptable-Risk systems are those which are exploitative, manipulative, or use subliminal techniques.

These AI systems are banned from sale on the EU Market.

Examples of prohibited AI use include:

- Cognitive behavioural manipulation of people or specific vulnerable groups: for example, voice-activated toys that encourage dangerous behaviour in children
- Social scoring: classifying people based on behaviour, socio-economic status or personal characteristics
- Biometric identification and categorisation of people
- Real-time and remote biometric identification systems, such as facial recognition

AI DEVELOPMENT LIFECYCLE



AI development lifecycle

Turning ethical principles, frameworks and guidelines into concrete actions

The significance of principles, frameworks, and guidelines in AI development is well recognized, yet applying them practically during the creation of AI solutions presents several **obstacles**:

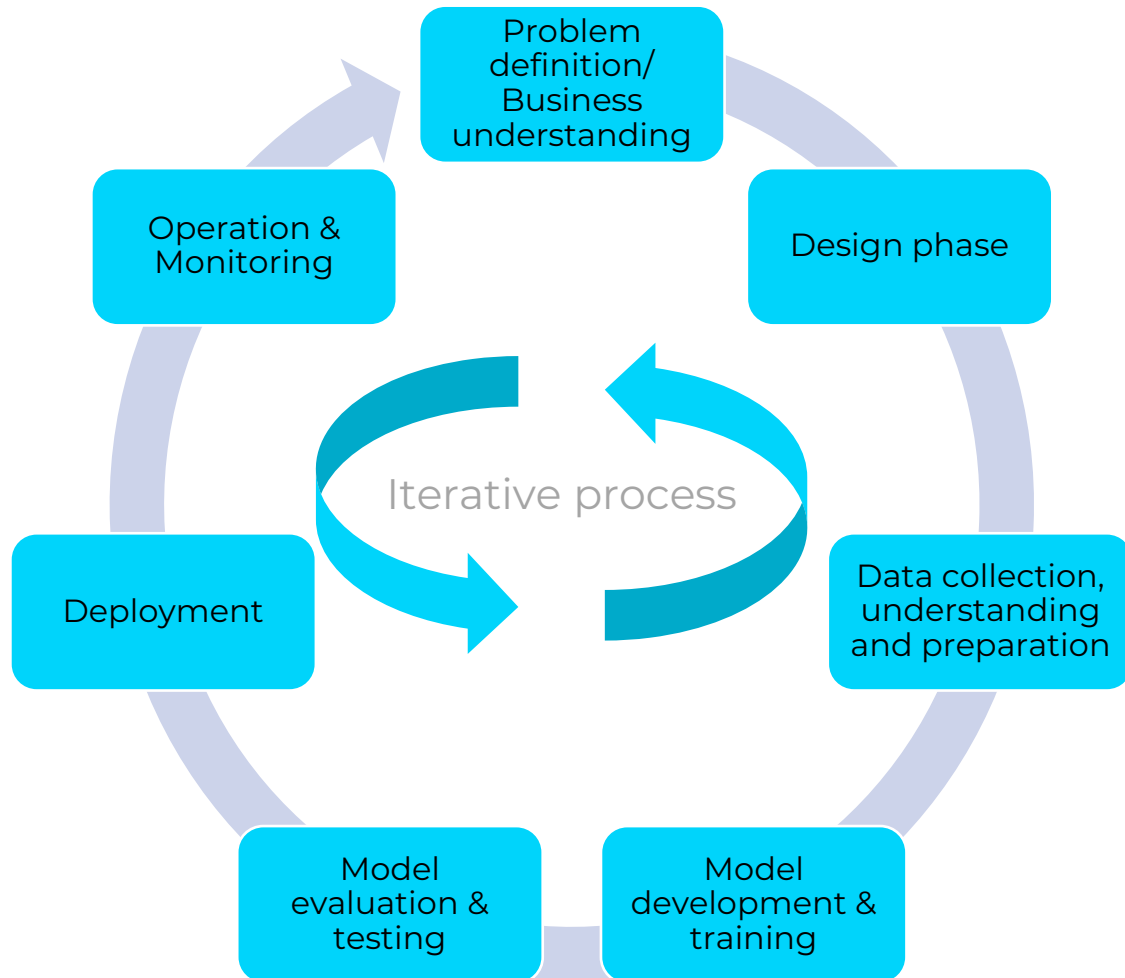
- AI principles tend to be abstract, diverse, and intricate.
- There's a complexity in converting nuanced human behavior into straightforward, universal tools.
- Principles are often subject to personal interpretation, influenced by individual backgrounds.
- There is an absence of universal standardization in regulations, frameworks, and guidelines (even if some generalization have been drawn as presented earlier)
- While frameworks are useful for issue identification, they may not provide actionable guidance on the specific steps to take

AI development lifecycle

Turning ethical principles, frameworks and guidelines into concrete actions: AI development lifecycle

- Available Tools designed to tackle ethical challenges in AI solutions are accessible, but they often cover only a select number of principles or AI solution development stages
- To transform ethical principles into concrete actions, it is recommended to view the AI development process as a sequence of steps starting with a business case and ending with the solution's operational mode. (AI Development Lifecycle)
- AI Development Lifecycle, adopts a holistic view of ethical considerations throughout all stages of development

AI development lifecycle

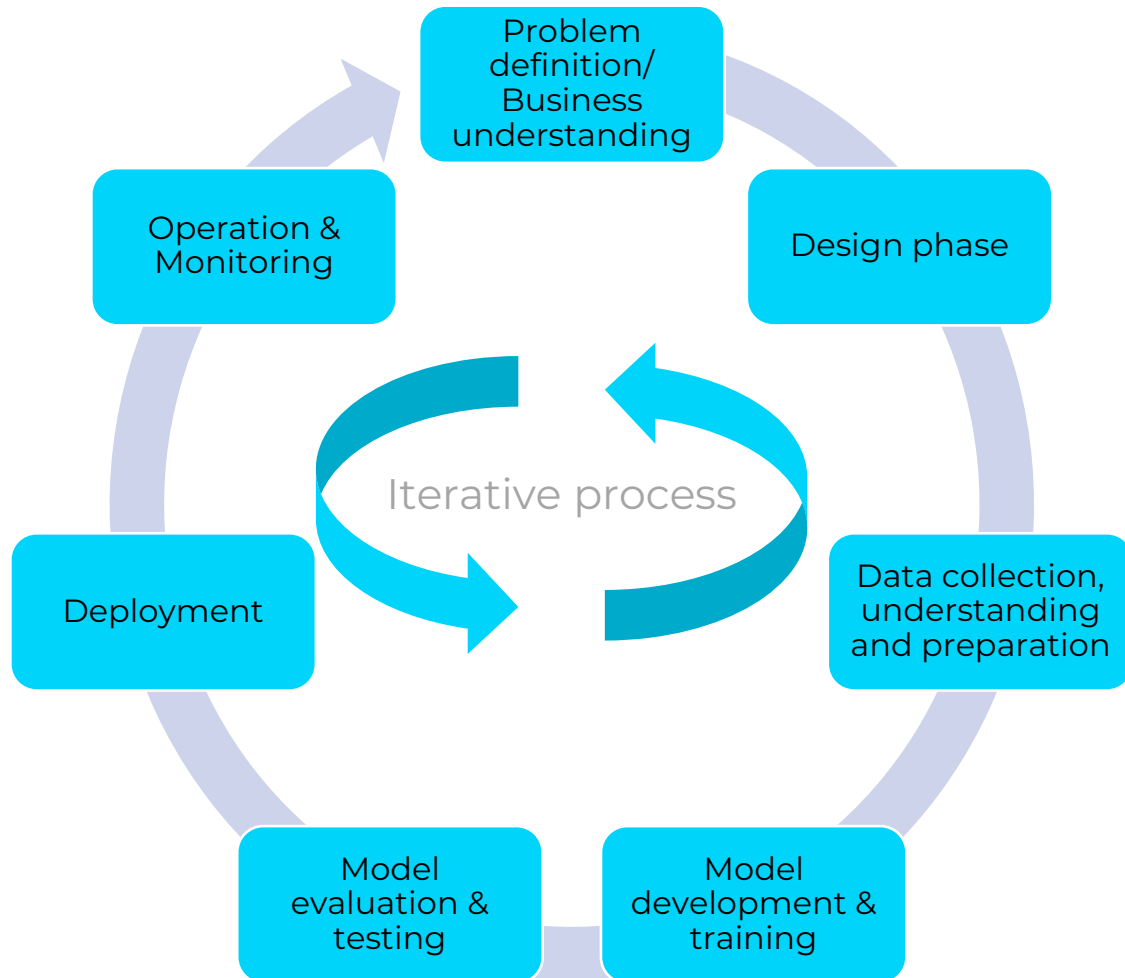


- Like any project, AI design project is recommended to follow design steps in order to transform innovative ideas into practical, well-defined solutions
- It's important to understand that each step plays a vital role in the development of the solution, forming an AI Development Lifecycle
- The process is ITERATIVE meaning that there could be a need to go back to the previous stage in case needed and modify the solution

AI Development design stages presented here are a combination of the different models described by the Data Science Process Alliance (2023), Morley et.al' (2019), Prehm's (2022) and Rochel et.al'a (2020) articles.

Data Science Process Alliance /Morley (2019) / Prem (2023) / Rochel & Evequoz (2020)

AI development lifecycle



- Like any project, AI design project is recommended to follow design steps in order to transform innovative ideas into practical, well-defined solutions
- It's important to understand that each step plays a vital role in the development of the solution, forming an AI Development Lifecycle
- The process is ITERATIVE meaning that there could be a need to go back to the previous stage in case needed and modify the solution

AI Development design stages presented here are a combination of the different models described by the Data Science Process Alliance (2023), Morley et.al' (2019), Prehm's (2022) and Rochel et.al'a (2020) articles.

Data Science Process Alliance /Morley (2019) / Prem (2023) / Rochel & Evequoz (2020)

AI development lifecycle



PROBLEM
DEFINITION AND
BUSINESS
UNDERSTANDING

Problem definition is crucial to understand customer needs and the role of the AI solution.



DESIGN PHASE

Design phase turns the business requirements into design requirements for engineers



DATA
COLLECTION
UNDERSTANDING
AND
PREPARATION

Data collection, understanding and preparation are a foundation for the solution



MODEL
DEVELOPMENT
AND TRAINING

Model development and evaluation are aimed to turn the data into a feasible solution addressing the initial problem



MODEL
EVALUATION &
TESTING

Model needs to be tested extensively and the functionality is mirrored into customer needs



DEPLOYMENT

Deployment is bringing the solution to its users



OPERATION
AND
MONITORING

The solution needs to be monitored during time and maintained according to changing needs

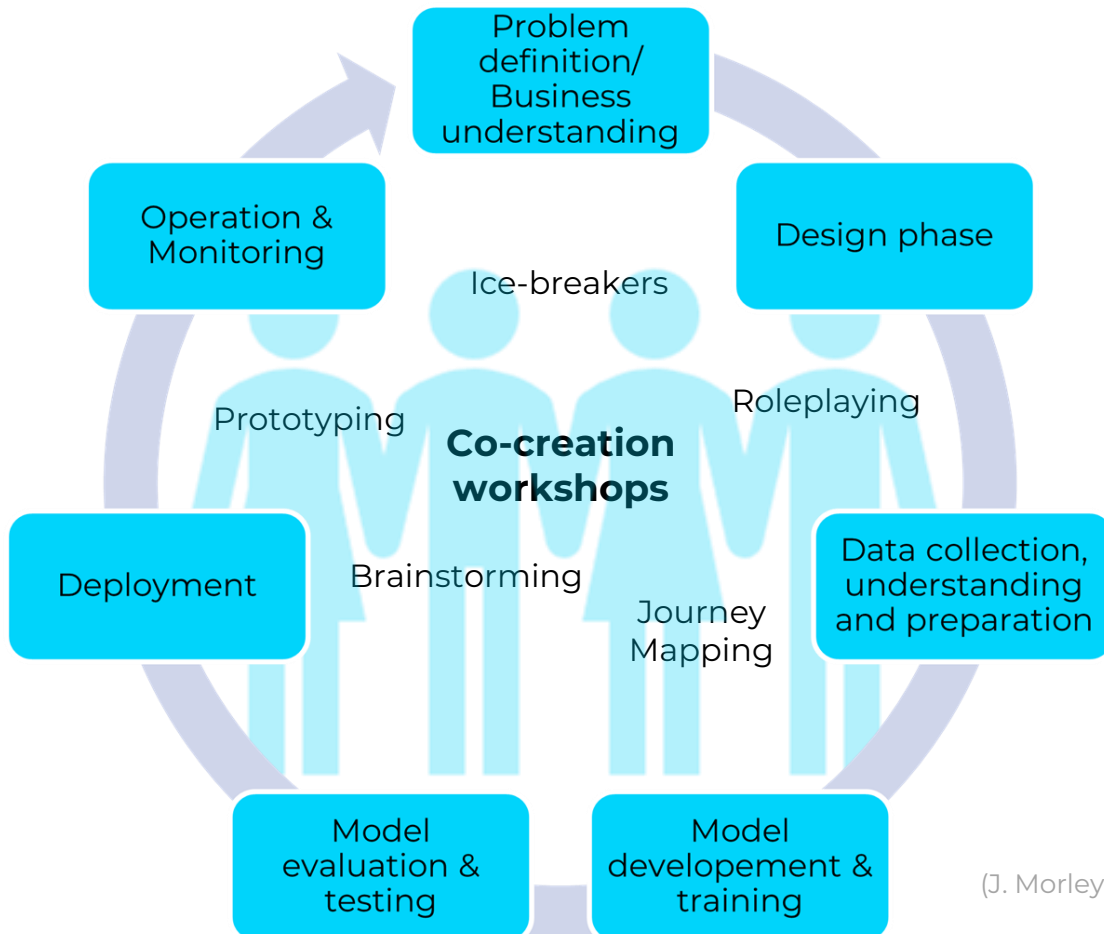
STAKEHOLDER ENGAGEMENT IN AI SOLUTION DEVELOPMENT



IMAGE SOURCE | Freepik

Stakeholders engagement in AI solution development

Why is it important to engage various stakeholders in AI solution development?

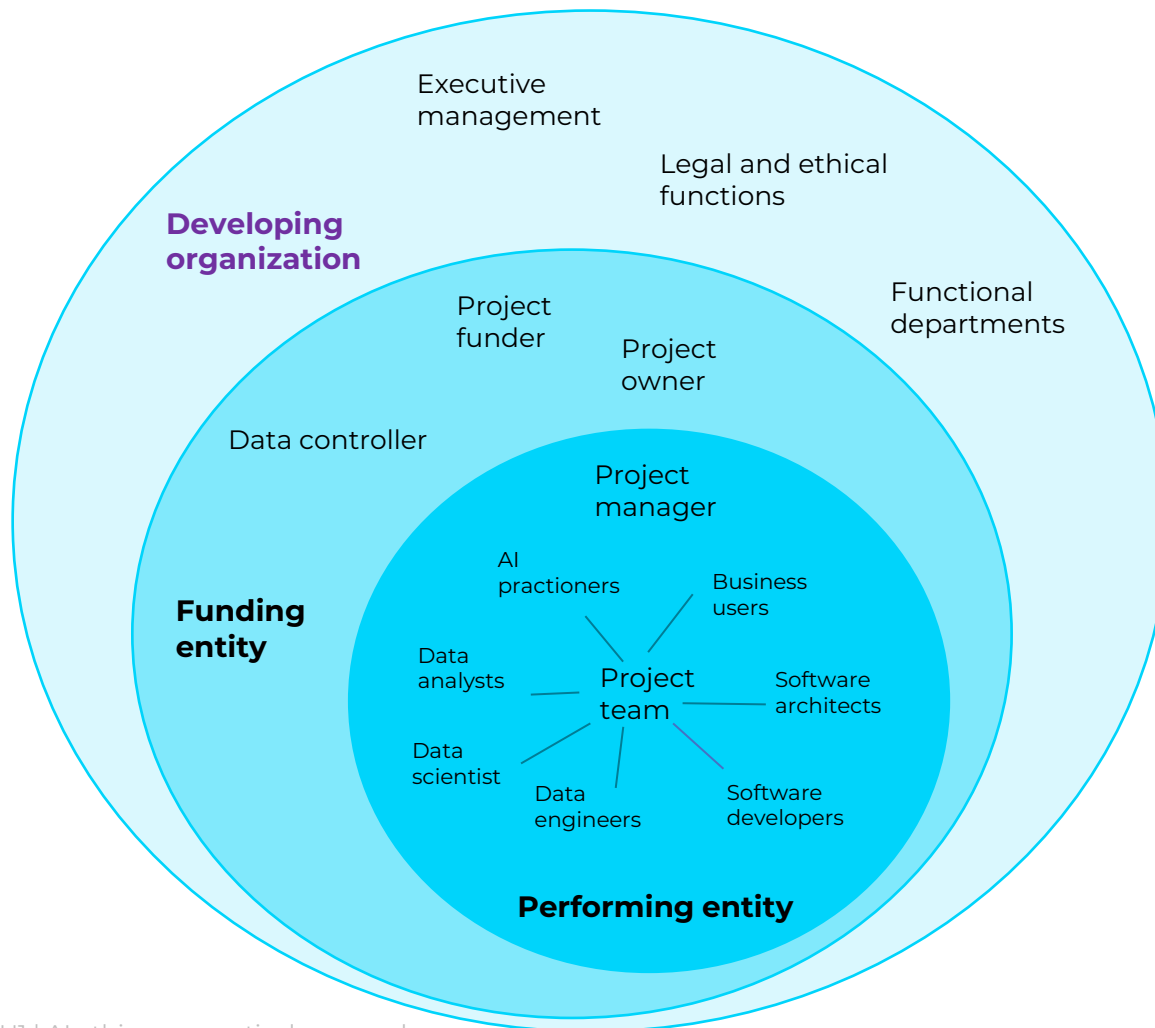


- To ensure AI development impacts are considered from different viewpoints and to develop trustworthy systems it is essential to engage and consult multiple stakeholders in the development process including those who will be affected by the AI solution.
- Co-creation is a practice used to collaborate with the stakeholders including end users or customers during a design process to share insights between the stakeholders with different roles and expertise and allows to better understand the needs of the end users and thus to design solutions that takes different users into account.

(J. Morley et al., 2019) / (UNESCO, 2023) / (Interaction Design Foundation, Co-creation) / (Anika Sanin, 2020)

Stakeholders engagement in AI solution development

Development stakeholders



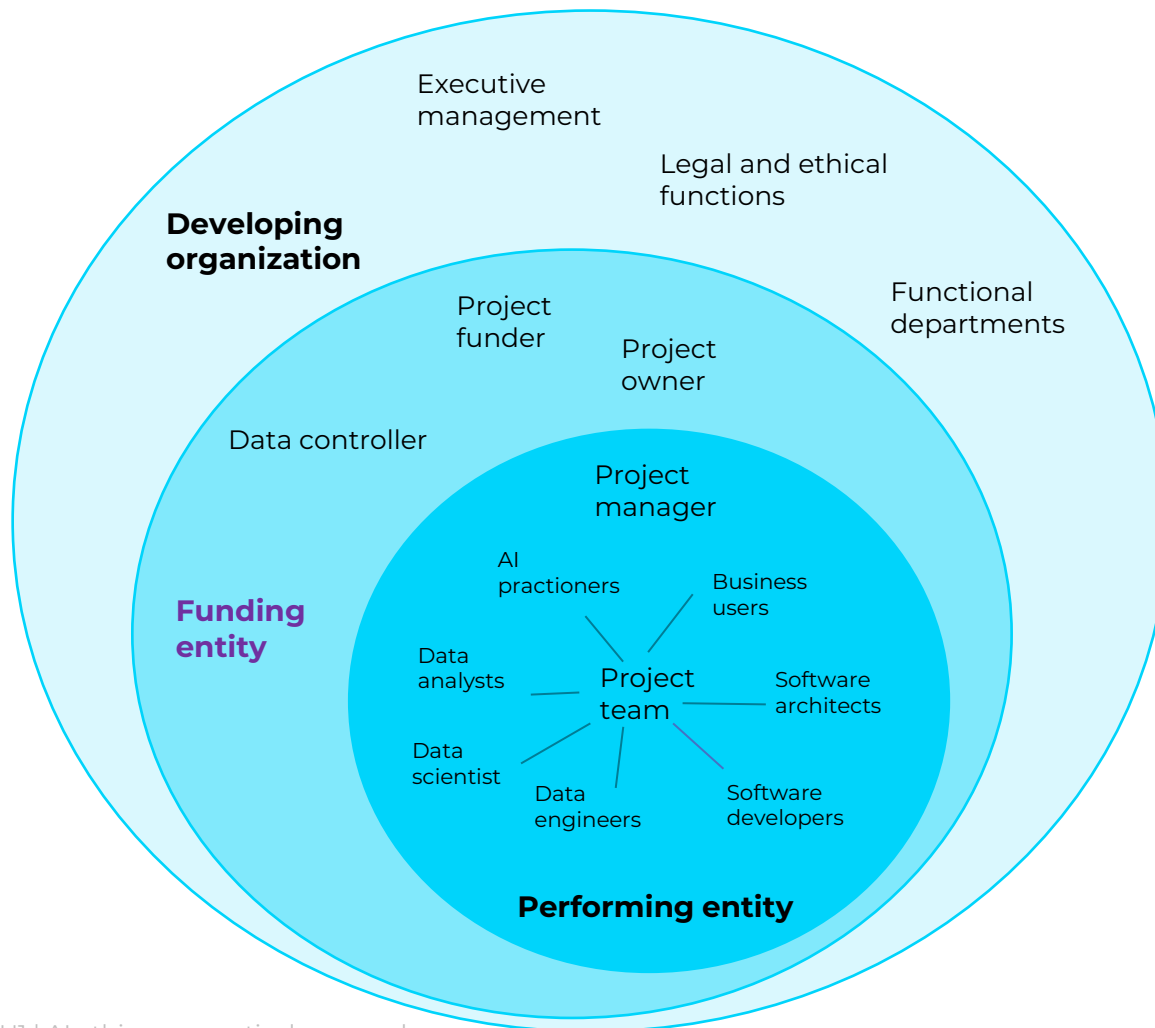
Developing organization

- Sponsors the AI project and determines the scope of the project.
- Project roles are **executive management, functional departments, and legal and ethical functions**:
 - **Executive management** is responsible for company's strategic, financial and developmental decisions.
 - **Functional department** is a specialized organizational unit within a company that focuses on a specific set of tasks or activities, often grouped based on common functions such as finance, marketing, or human resources.
 - **Legal and ethical functions** within an organization involve various roles and responsibilities to ensure compliance with laws, regulations, and ethical standards.

(G.J. Miller, 2022)

Stakeholders engagement in AI solution development

Development stakeholders



Funding entity

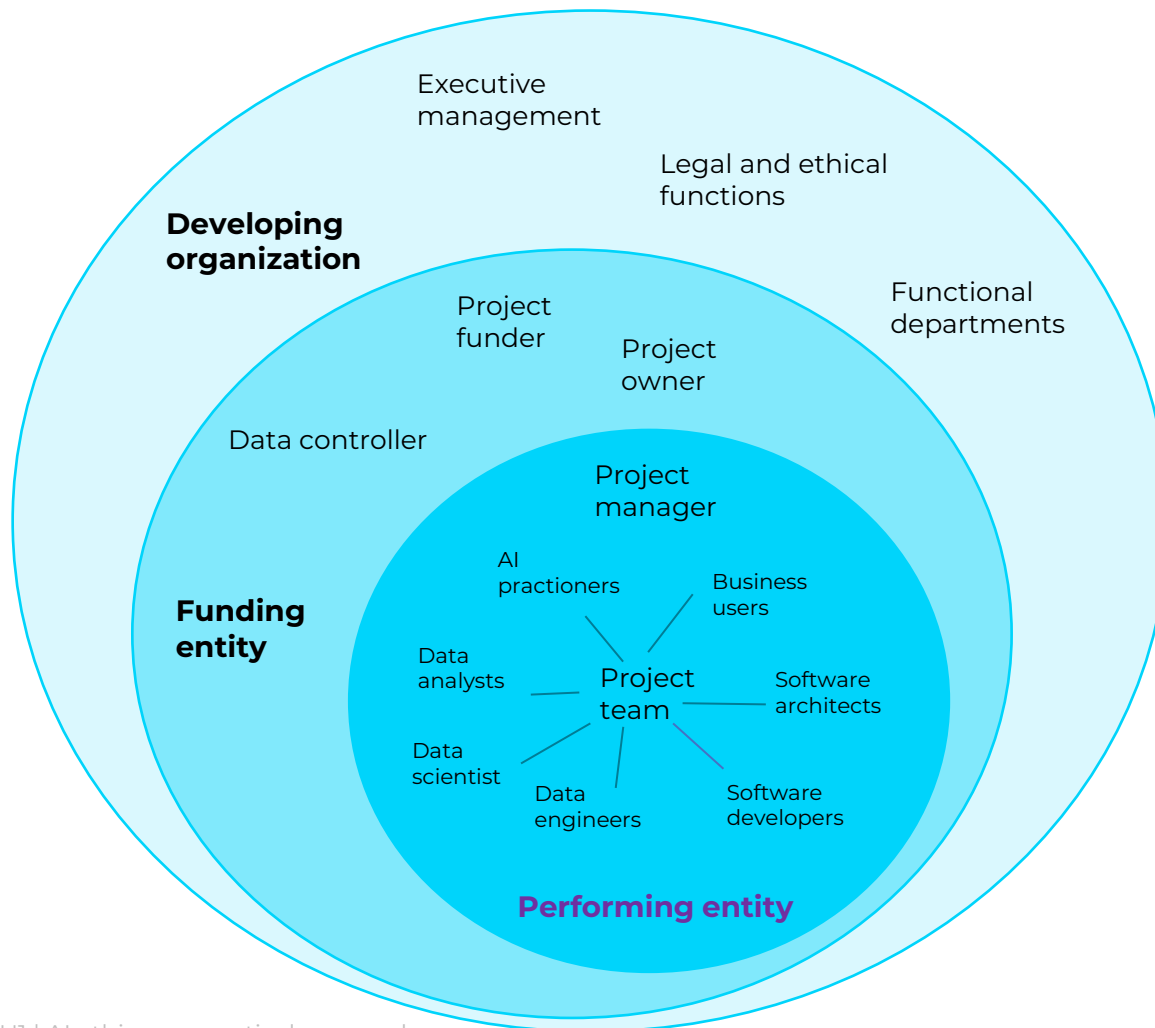
Project roles are **project funder**, **project owner** and **data controller**.

- **Project funder** funds the project and allocates the resources for the project.
- **Project owner** offers strategic direction.
- **Data controller** is responsible for the management of the data and its use in system development.

(G.J. Miller, 2022)

Stakeholders engagement in AI solution development

Development stakeholders



Performing entity

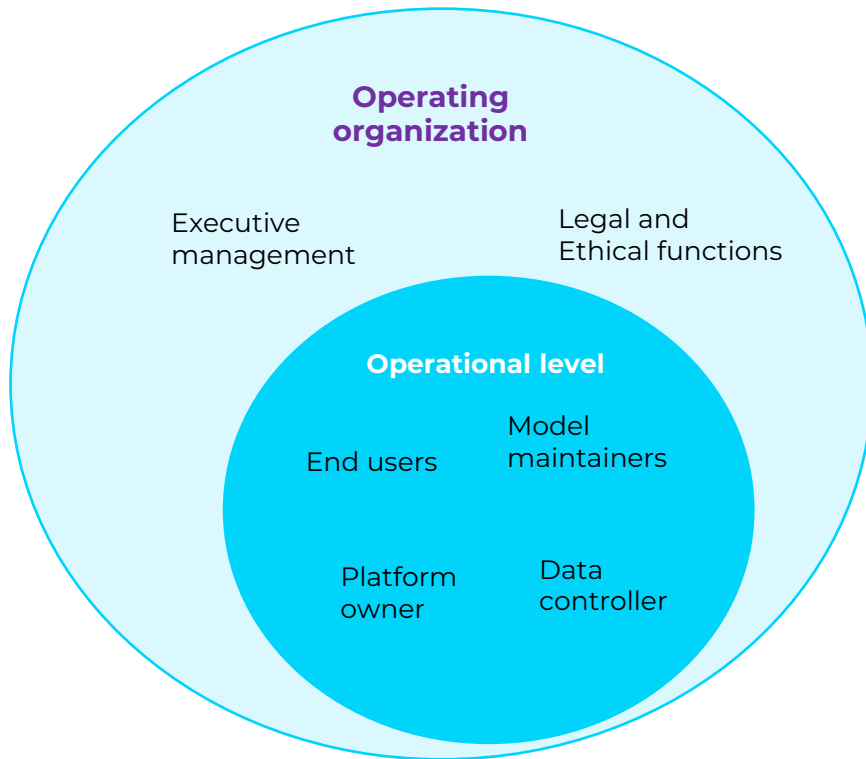
Project roles are **project managers** and **project team**.

- **Project manager** is responsible for the project's outcomes and ensures that ethical, privacy and security standards and expectations are respected and that relevant stakeholders are involved.
- **AI project team** consists of software architects, software developers, data engineers, data scientists, data analysts, AI practitioners and business users.

(G.J. Miller, 2022)

Stakeholders engagement in AI solution development

Usage stakeholders

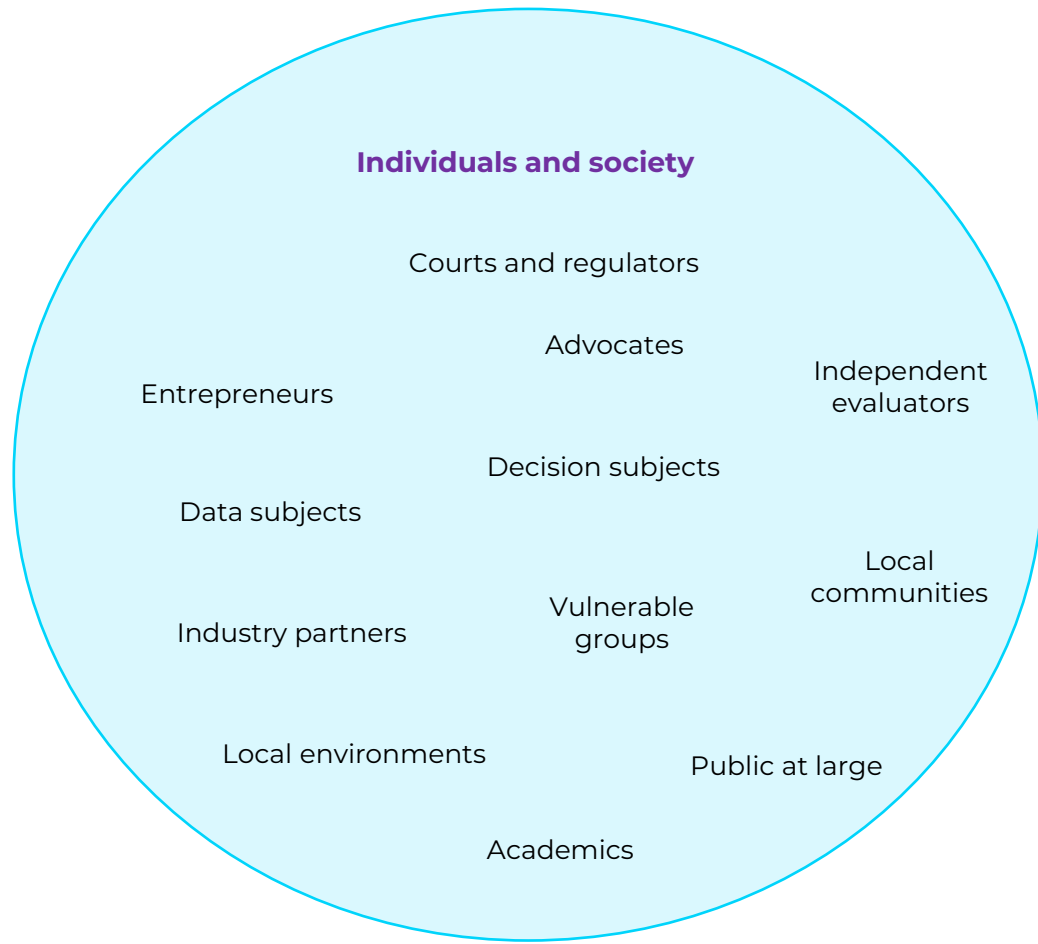


Operating organization

- An operating organization is a client stakeholder.
- As well as in the development organization, the operating organization's project roles include **executive management** and **legal and ethical functions**.
- On the operations level the project roles are **end users**, **model maintainers**, **data controllers** and **platform owners**.
 - **End users** are individuals or groups who interact with the AI system through work or service contracts with the operating organization or as consumers.
 - **Model maintainers** are responsible for the ongoing management, updates, and optimization of the machine learning models used in the AI system.
 - **Data controllers** determine the purposes and means of processing personal data within the AI system.
 - **Platform owners** are responsible for the overall infrastructure and deployment of the AI system.

Stakeholders engagement in AI solution development

External stakeholders



Individuals and society

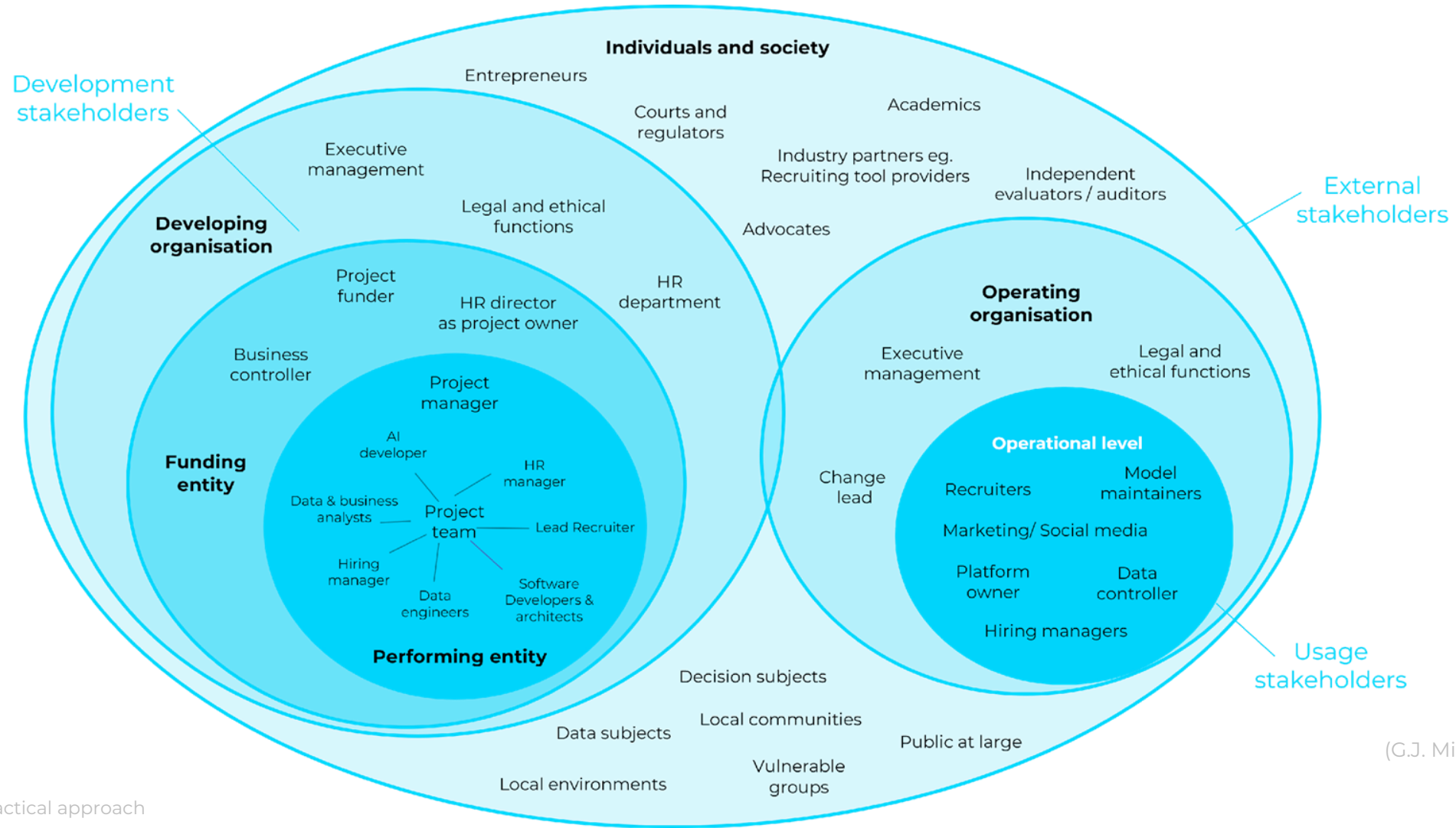
- **Individual subjects** are people who have agreements with the development or operating organization, like service or employment contracts. This group includes data subjects, decision subjects, and workers, and they are considered passive stakeholders.
- **Society** consists of individuals and the public. They might be negatively affected by the development or use of the AI system.
- This group comprises local communities, environments, and vulnerable groups like people with disabilities, minority groups, minors, and the general public.

AI engineers are part of a greater societal discussion, where they need to play an active role in integrating other perspectives.

- **Representatives** are groups and organizations that are not affected by the AI system but might represent other in the project.
- **Formal representatives** include courts, laws, regulators, and labor unions.
- **Entrepreneurs** can be involved in the development of the AI system.
- **Independent evaluators** are people or groups from outside, for example, reporters, auditors, or reviewers.

Stakeholders engagement in AI solution development

Stakeholders in AI development



ETHICAL PRINCIPLES AND STAKEHOLDER ENGAGEMENT IN AI DEVELOPMENT LIFECYCLE



Ethical Principles and stakeholders in AI development lifecycle



Throughout the AI lifecycle, it's important to maintain a focus on ethical considerations and stakeholder engagement.

The following pages go through the lifecycle from three perspectives:

- 1. Stage description**
- 2. Main ethical principles**
- 3. Involved stakeholders**

Not all ethical principles hold the same weight in every context or at every stage of the development and thus only the main considerations and viewpoints are presented in the following.

Ethical Principles and stakeholders in AI development lifecycle

Problem definition & business understanding



Crucial first stage for a successful project

- Understand the problem to be solved or the opportunity to be investigated.
- Conduct research and stakeholder interviews to formulate a clear problem statement.
- Understand customer needs and how they translate into a viable business solution.
- Define the goals and success criteria of the solution.
- Conduct a feasibility study to determine if AI is the most appropriate solution.
- Assess both the technical and ethical feasibility of the project

Ethical Principles and stakeholders in AI development lifecycle

Problem definition & business understanding

Ethical principles to be considered



Setting the foundation - address basic ethical issues such as human rights, environmental impact, and societal concerns.

Beneficence - ensure the AI solution enhances individual well-being and societal good. Key areas to consider:

- **Stakeholder Participation**
 - Build trustworthy systems that support human flourishing.
 - Involve impacted individuals in the development process.
- **Protection of Fundamental Rights**
 - Safeguard individual rights, reinforcing ethical integrity.
- **Environmental Sustainability**
 - Minimize environmental impact through sustainable practices.
- **Justification**
 - Clearly define and link the system's purpose to tangible societal benefits.

Morley (2019) / Prem (2023)

Ethical Principles and stakeholders in AI development lifecycle

Problem definition & business understanding

Ethical principles to be considered



Non-Maleficence

- Ensure that the problem identification and proposed solution do not harm individuals or communities.
- Carefully consider and monitor the system's impact on physical and mental well-being.
- Assess potential safety risks and strategies for mitigation.

Justice

- Evaluate the system's impact on institutions, democracy, and society.
- Consider whether certain groups are favored or disadvantaged.
- Examine if the project might infringe on individual autonomy.

Ethical Principles and stakeholders in AI development lifecycle

Problem definition & business understanding Stakeholders



- **End-users:** primary users of the AI solution
- **AI/data scientists:** formulate the problem and conceptualize a solution
- **Communities** affected
- **Domain experts:** deep knowledge of industry where AI is being applied
- **AI developers:** early involvement brings them into deep understanding of the problem
- **Project managers:** overseeing the whole progress of the project
- **Business stakeholders:** ensuring alignment with business objectives and resources
- **Ethicists and social scientists:** highlighting potential impacts

(D. De Silva & D. Alahakoon, 2022)

Ethical Principles and stakeholders in AI development lifecycle

Design phase



This stage transforms the groundwork laid during the problem definition stage into detailed plans for building and deploying the AI system

- Specify what the project aims to achieve (goals and outcomes)
- Evaluate the financial cost against the anticipated benefits.
- Outline the functionalities that the AI system must possess.
- Decide on the approach for data utilization, such as the use of pre-trained models.
- Establish metrics to measure the project's success.
- Integrate ethical guidelines to govern the development process.
- Develop a timeline and manage stakeholder involvement.

Ethical Principles and stakeholders in AI development lifecycle

Design phase

Ethical principles to be considered



Justice | Avoiding Unfair Bias

- **Design requirements**
 - Implement mechanisms to identify and mitigate biases in data and algorithms.
 - Aim for equitable outcomes across different groups.
- **Developer awareness**

Ensure developers are conscious of potential biases, particularly related to e.g. race, color, descent, gender, age, language, religion, political opinion, national, ethnic, social origin, economic status, condition of birth, disability
- **Communication and response**

Establish clear protocols to communicate, report, and respond to biases, negative impacts, or trade-offs.
- **Data management**

Consider data storage systems to prevent siloed data bias.
- **Risk management**

Develop "What happens if" measures to handle unexpected outcomes.

Ethical Principles and stakeholders in AI development lifecycle

Design phase

Ethical principles to be considered



Explicability / Transparency

- Ensure clear documentation and user-friendly explanations of AI decisions.
- Ensure transparency of data sources and algorithmic processes.

Beneficence

Engage stakeholders in the development process to consider diverse ethical aspects.

Autonomy

Implement human oversight mechanisms to maintain human values in AI solutions.

Ethical Principles and stakeholders in AI development lifecycle

Design phase Stakeholders



- **Software architects:** lay the technical foundation for the AI solution, considering both current and future needs.
- **User Experience (UX) designers:** critical for creating interfaces that enhance user adoption and satisfaction and for ensuring that users can interact effectively with the AI system.
- **Ethics experts:** address ethical concerns.
- **AI practitioners:** collaborating with UX designers to integrate AI components seamlessly.
- **Legal and ethical functions:** mitigate legal risks by addressing compliance requirements early in the design phase and help in designing solutions that respect data protection and privacy regulations.
- **People and communities affected by the AI model:** participatory design approach.

Ethical Principles and stakeholders in AI development lifecycle

Data collection, understanding and preparation



Once the goal and the plan to achieve it have been established, the next step involves working with relevant data.

- Gather all necessary data for the project.
- Describe, explore, verify, select, clean, construct, integrate, and format the data into suitable categories and concepts for the intended purpose.
- Address and resolve issues related to missing data to ensure completeness.
- Recognize and understand any assumptions underlying the existing data.
- Clarify how datasets will be utilized for decision-making.
- Maintain high data quality, especially in cases of unclear or missing information.

This is a time-consuming stage where the foundation of the working solution is built. The reliability and performance of the solution are directly correlated with the quality of the underlying data.

Ethical Principles and stakeholders in AI development lifecycle

Data collection, understanding and preparation

Ethical principles to be considered



Non-maleficence 1/2

- **Data integrity and quality causing bias**

- Address potential biases favoring certain demographics. Ensure accuracy when combining datasets (e.g. persons with same name should not be mixed)
- Conduct thorough due diligence to assess and verify data accuracy, ensuring that it does not perpetuate unfair biases.
- Avoid using siloed data based on specific demographics, which can enhance biases.
- Implement a unified data repository, such as a data warehouse, to minimize bias (note: expensive).



Can bias be a relevant factor in the AI solution? E.g., car insurance influenced by types of accidents and demographics involved. Legal constraints: Gender cannot determine car insurance rates, as per the European Court of Justice, 2012.

Ethical Principles and stakeholders in AI development lifecycle

Data collection, understanding and preparation Principles



Non-maleficence 2/2

- **Data Privacy and Confidentiality:**
 - Address typical risks around data privacy and confidentiality during data collection.
 - Ensure AI systems uphold data privacy throughout the lifecycle, adhering to regulations like GDPR.
- **Resilience to Attack and Security:**
 - Protect data sets from attacks and address potential security risks to maintain robust data security.

Ethical Principles and stakeholders in AI development lifecycle

Data collection, understanding and preparation Principles



- **Data scientists** find patterns and trends in data to turn it into knowledge.
- **AI engineers:** responsible for assessing and verifying the data.
- **Data analysts** contribute to understanding the characteristics of the data, helping to inform data preparation strategies.
- **Data engineers:** responsible for preparing the data and ensuring its quality.
- **Data protection officer (DPO):** responsibility to make sure their organization follows the rules when handling personal data of employees, customers, providers, or other individuals (referred to as data subjects).
- **Ethicists:** address ethical considerations in the handling and selection of data.

Ethical Principles and stakeholders in AI development lifecycle

Model development and training



In this phase, an AI model is developed to address the defined problem. This includes several critical steps:

- Select and configure the algorithmic model and tools, defining their use upfront.
- List and evaluate the implications of the chosen tools, especially considering potential drawbacks and trade-offs.
- Decide which elements of the model should be emphasized and which may become secondary or even hidden.
- Engage in multiple rounds of refinement during the training process, continually improving the model.
- Selection of tools often involves balancing conflicting interests. For example, deciding whether it is preferable to risk having spam emails in the actual inbox or valid emails ending up in the spam folder.

This stage is crucial as it directly influences the effectiveness and efficiency of the AI solution.

Ethical Principles and stakeholders in AI development lifecycle

Model development and training

Ethical principles to be considered



Explicability

- While focusing on technical aspects, it's crucial to check for explicability and whether biases exist within the AI system, offering a prime opportunity for intervention.
- Maintain transparency throughout the development process to ensure stakeholders can understand it. This involves clear documentation of data sources, model choices, and the rationale behind these decisions.

Justice

- Recognize and transparently report any trade-offs between efficiency and other requirements.
- Establish a process to acknowledge and evaluate the impact of these trade-offs.
- AI engineers must justify their decisions, taking into account the potential negative impacts on affected individuals.

Morley (2019) Prem (2023)

Ethical Principles and stakeholders in AI development lifecycle

Model Development and Training Stakeholders



- **AI/ML scientists** transform the problem definition into a prototypical AI model.
- **AI engineers** have expertise to select the most suitable algorithmic tools for certain purposes.
- **Data scientists** are central to the development of machine learning models and their optimization.
- **Ethics experts** are essential for integrating ethical principles into the development process to avoid unintended consequences.
- **Business users** ensure that the models meet the business requirements and expectations.

(D. De Silva & D. Alahakoon, 2022), (J. Rochel & F. Évequoz, 2020)

Ethical Principles and stakeholders in AI development lifecycle

Model evaluation & testing



Once the model has been developed and trained, its performance needs to be tested and evaluated:

- Evaluate the model against pre-defined objectives and success criteria.
- Clearly articulate the underlying assumptions and standards upon which the datasets are based.
- Conduct tests using unseen data or steps in the process to ensure robustness.
- If the performance is unsatisfactory, consider altering model parameters, architecture, or even the datasets behind the data model.
- Any shortcomings should be openly and honestly discussed with the project leader and other stakeholders, as they might significantly impact individuals.
- Agree on whether to proceed with more iterations or move on to the deployment phase.

This stage is crucial for ensuring the model's reliability and effectiveness before it goes into production.

Ethical Principles and stakeholders in AI development lifecycle

Model evaluation & testing

Ethical principles to be considered



Justice / Beneficence / Non-maleficence

- **Ethical Audits and Metrics:**

- Reassure the adherence to ethical principles such as justice, beneficence, and non-maleficence through audits and metrics designed for auditability.

- **Security and Resilience:**

- Test resilience to attacks and establish a fallback plan to ensure system security.

- **Addressing Negative Impacts & possible risks:**

- Implement strategies for minimizing and reporting negative issues.
- Establish a redress mechanism to address potential harms if things go wrong.
- Formulate risk assessment factors focusing on privacy, social implications, and bias.

- **Explicability:**

- Clearly explain both the technical processes and the human decisions involved in the model development.

Be aware that engineers under pressure to finalize solutions for deployment may face increased risk of overlooking critical issues.

Ethical Principles and stakeholders in AI development lifecycle

Model evaluation & testing Stakeholders



- **AI engineers** interpret the results after modelling.
- **Quality Assurance (QA) teams** ensure the models meet quality standards and are free of errors or unintended consequences.
- **Ethics experts** ensure the ethical use of AI models.
- **User advocates/Representatives** ensure that the AI models align with user expectations and needs.
- **Legal and Compliance teams** mitigate legal risks by ensuring compliance with relevant regulations.

(J. Rochel & F. Évequoz, 2020) / (Moore, 2023)

Ethical Principles and stakeholders in AI development lifecycle

Deployment



Once the AI solution has been finalized to satisfaction, it is deployed to a production environment to address real-world problems. The complexity of this phase depends on the solution itself.

- Initially, the solution is typically introduced to a smaller group of experts and users.
- Either integrate the model with existing systems, create an application or service that utilizes the model or leverage the insights in an offline context, such as a report to management.
- Establish a feedback loop for monitoring performance and outcomes, with subsequent iterations if necessary.
- Implement risk management strategies to address potential issues during deployment.
- Plan for ongoing monitoring and maintenance in the operational phase.
- Review the entire project and produce a final report documenting the process and outcomes.

This structured approach ensures the AI solution is effectively integrated and capable of performing as intended in real-world scenarios.

Ethical Principles and stakeholders in AI development lifecycle

Deployment

Ethical principles to be considered



Autonomy

- **Human agency**

- Ensure users are aware when decisions, content, advice, or outcomes are generated by an AI. Highlight the interaction with the AI solution to prevent over-reliance.

- **Human oversight**

- Guarantee that human oversight is in place to intervene when necessary, ensuring decisions can be reviewed or altered by humans.

Justice

- Implement strong auditing measures to regularly assess system performance and ethical adherence.
- Establish mechanisms for users to appeal decisions or file complaints, ensuring protections are in place for whistleblowers.

Ethical Principles and stakeholders in AI development lifecycle

Deployment

Ethical principles to be considered



Explicability

- Highlight the importance of explaining both the technical processes and related human decisions to users.
- Tackle the challenge of simplifying complex human behaviors into tools that are easily understandable, avoiding technical jargon -> this increases trust and system acceptance
- Maintain transparency about AI functionalities and limitations.

Evaluating long-term impacts

- Assess the longer-term impacts on society and the environment at this stage.
- Consider necessary improvements to ensure the AI system contributes positively over time.

Ethical Principles and stakeholders in AI development lifecycle

Deployment Stakeholders



- **AI/ML engineers** transform the prototypical AI model into a deployed service or solution that is accessible for all stakeholders and end-users.
- **Security experts** safeguard the deployed system against cybersecurity risks.
- **End users** to identify any issues or limitations that may not have been apparent during testing in controlled environments.
- **User advocates/representatives** collect user feedback on the usability and performance of the deployed AI solution and ensure that the deployed solution meets user expectations and needs.

(D. De Silva & D. Alahakoon, 2022) / (Junklewitz et al., 2023)

Ethical Principles and stakeholders in AI development lifecycle

Operation and monitoring



After deployment, the AI solution requires ongoing maintenance, updates, and continuous evaluation.

- Ensure that the model continues to perform as expected by regularly monitoring its effectiveness.
- Continuously monitor end-user activity to gather operational insights and detect potential issues.
- Regularly update the model with the newest data to maintain its accuracy and relevance.
- Refine the model based on user feedback to enhance functionality and performance.
- Iterative updates are often necessary and should be viewed as a normal part of the AI lifecycle, not as a failure.
- Monitor return on investment and other relevant metrics based on the initially set objectives.

This phase is critical to ensure that the AI solution remains effective and continues to meet the needs of its users.

Ethical Principles and stakeholders in AI development lifecycle

Operation and monitoring Ethical principles to be considered



Ethical principles monitoring

- Regularly revisit and assess the AI system's adherence to ethical principles, considering possible changes in datasets and model structures.

Continued stakeholder engagement

- Maintain active engagement with users, organizations, and broader society to gather ongoing feedback.
- Establish structured mechanisms to ensure regular and systematic collection of insights and responses.

Ethical Principles and stakeholders in AI development lifecycle

Operation and monitoring Stakeholders



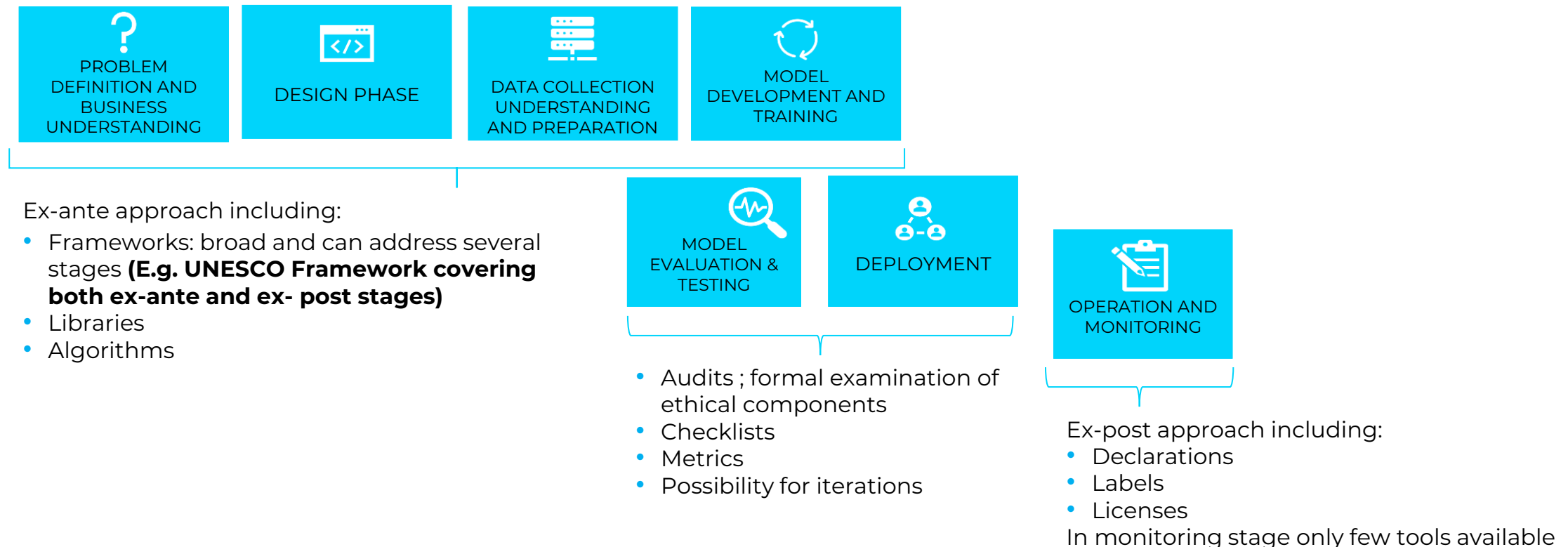
- **DevOps and IT Operations teams** monitor system performance and ensure continuous availability.
- **Safety certifiers** assess the security of the developed AI solution.
- **Expert panels, steering committees, or regulatory bodies** review the project in technical and ethical aspects.

(D. De Silva & D. Alahakoon, 2022) / (Immersant Data Solutions, 2023) / (Miller, 2022)

Ethical Principles and stakeholders in AI development lifecycle

AI development lifecycle Tools

Available tools for ethical considerations vary a lot based on the AI development stage. Many approaches can be valid in all steps, but some only in certain development steps. Below categorization is a rough picture of how the available tools can be divided between the design stages.



THANK YOU

Project number: 2022-1-ES01-KA220-HED-000085257



The European Commission's support for the production of this publication does not constitute of the contents, which reflect the views only of the authors , and the Commission cannot be held responsible for any use which may be made of the information contained therein.

