EU AI principles

Eugenio Fontenla Suárez

AI Act

- **Development vs. public interests**
  i.e health and safety, fundamental rights, democracy, rule of law, environmental protection.

- Benefit principle of **free movement of goods and services**
  Protecting fundamental rights while supporting innovative solutions.
  Public and private actors: SMEs including startups.

- **Human-centric approach** to AI and being a global leader in the development of secure, trustworthy and ethical artificial intelligence as stated by the European Council, protection of ethical principles, requested by the European Parliament.

- The regulation should **support innovation**, respect **freedom of science**, and should not undermine research and development activity.

  - AI systems and models specifically developed and put into service for the sole purpose of scientific research and development excluded + not limited prior to being placed on the market or put into service. As regards product oriented research, testing and development activity.

    - Still obligation to comply with this Regulation when an AI system is placed on the market or put into service as a result of such research and development activity and to the application of provisions on regulatory **sandboxes** and **testing in real world conditions**.

    - **Any other AI system** that may be used for the conduct of any research and development activity should remain subject to the provisions of this Regulation.

    - Under all circumstances, any research and development activity should be carried out in accordance with recognised **ethical and professional standards** for scientific research and should be conducted according to applicable Union law.

HLEG Guidelines

The 4 binding principles

There are **4 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.

They are specified as ethical **imperatives**, such that AI practitioners should always strive to adhere to them. Without imposing a hierarchy, we list the principles here below in a manner that mirrors the order of appearance of the fundamental rights upon which they are based in the EU Charter.

(i) The principle of respect for human autonomy

The fundamental rights upon which the EU is founded are directed towards ensuring respect for the freedom and autonomy of human beings.

Humans interacting with AI systems must be able to keep full and effective self determination over themselves, and be able to partake in the democratic process. AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans.

Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. This means securing human oversight over work processes in AI systems.

AI systems may also fundamentally change the work sphere. It should support humans in the working environment, and aim for the creation of meaningful work.

(ii) The principle of prevention of harm

AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. AI systems and the environments in which they operate must be safe and secure.

They must be technically robust and it should be ensured that they are not open to malicious use. Vulnerable persons should receive greater attention and be included in the development, deployment and use of AI systems.

Particular attention must also be paid to situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and consumers or governments and citizens.

Preventing harm also entails consideration of the natural environment and all living beings.

(iii)The principle of fairness

The development, deployment and use of AI systems must be fair.

While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension.

The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation. If unfair biases can be avoided, AI systems could even increase societal fairness.

Equal opportunity in terms of access to education, goods, services and technology should also be fostered. Moreover, the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice.

Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives. The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them.In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.

(iv) **The principle of explicability**

Explicability is crucial for building and maintaining users' trust in AI systems.

This means that processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. Without such information, a decision cannot be duly contested.

An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights.

The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.

These requirements are applicable to different stakeholders partaking in AI systems' life cycle: developers, deployers and end-users, as well as the broader society.

By developers, we refer to those who research, design and/or develop AI systems.

By deployers, we refer to public or private organisations that use AI systems within their business processes and to offer products and services to others.

End-users are those engaging with the AI system, directly or indirectly. Finally, the broader society encompasses all others that are directly or indirectly affected by AI systems.

Different groups of stakeholders have different roles to play in ensuring that the requirements are met:

- **Developers** should implement and apply the requirements to design and development processes;

Developers hold the primary responsibility for ensuring that ethical considerations are integrated into the very fabric of language model development. Embracing the principle of "ethical by design," they must proactively address biases, privacy concerns, and potential misuse during the creation phase. This involves meticulous dataset curation, algorithm design, and ongoing evaluation of the model's outputs.

- **Deployers** should ensure that the systems they use and the products and services they offer meet the requirements;

- **End-users** and the broader society should be informed about these requirements and able to request that they are upheld.

The 7 non-binding principles

HLEG developed seven non-binding ethical principles for AI which should help ensure that AI is trustworthy and ethically sound.

1. Human agency and oversight

Means that AI systems are developed and used as a tool that serves people, respects human dignity and personal autonomy, and that is functioning in a way that can be appropriately controlled and overseen by humans.

○ Including fundamental rights, human agency and human oversight.

AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights, and allow for human oversight.

- **Fundamental rights**

Like many technologies, AI systems can equally enable and hamper fundamental rights. They can benefit people for instance by helping them track their personal data, or by increasing the accessibility of education, hence supporting their right to education. However, given the reach and capacity of AI systems, they can also negatively affect fundamental rights. In situations where such risks exist, a fundamental rights impact assessment should be undertaken.

This should be done prior to the system's development and include an evaluation of whether those risks can be reduced or justified as necessary in a democratic society in order to respect the rights and freedoms of others. Moreover, mechanisms should be put into place to receive external feedback regarding AI systems that potentially infringe on fundamental rights.

- **Human agency**

Users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system. AI systems should support individuals in making better, more informed choices in accordance with their goals.

AI systems can sometimes be deployed to shape and influence human behaviour through mechanisms that may be difficult to detect, since they may harness sub-conscious processes, including various forms of unfair manipulation, deception, herding and conditioning, all of which may threaten individual autonomy. The overall principle of user autonomy must be central to the system's functionality. Key to this is the right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them.

- **Human oversight**

Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as a human-in-theloop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach.

HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable.

HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation.

HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation.

This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system. Moreover, it must be ensured that public enforcers have the ability to exercise oversight in line with their mandate.

Oversight mechanisms can be required in varying degrees to support other safety and control measures, depending on the AI system's application area and potential risk. All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required.

2. Technical robustness and safety

Means that AI systems are developed and used in a way that allows robustness in case of problems and resilience against attempts to alter the use or performance of the AI system so as to allow unlawful use by third parties, and minimise unintended harm.

○ Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility.

A crucial component of achieving Trustworthy AI is technical robustness, which is closely linked to the principle of prevention of harm.

Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. In addition, the physical and mental integrity of humans should be ensured.

- **Resilience to attack and security**

AI systems, like all software systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries, e.g. hacking. Attacks may target the data (data poisoning), the model (model leakage) or the underlying infrastructure, both software and hardware. If an AI system is attacked, e.g. in adversarial attacks, the data as well as system behaviour can be changed, leading the system to make different decisions, or causing it to shut down altogether.

Systems and data can also become corrupted by malicious intention or by exposure to unexpected situations. Insufficient security processes can also result in erroneous decisions or even physical harm.

For AI systems to be considered secure, possible unintended applications of the AI system (e.g. dual-use applications) and potential abuse of the system by malicious actors should be taken into account, and steps should be taken to prevent and mitigate these.

- **Fallback plan and general safety**

AI systems should have safeguards that enable a fallback plan in case of problems. This can mean that AI systems switch from a statistical to rule-based procedure, or that they ask for a human operator before continuing their action.

It must be ensured that the system will do what it is supposed to do without harming living beings or the environment. This includes the minimisation of unintended consequences and errors.

In addition, processes to clarify and assess potential risks associated with the use of AI systems, across various application areas, should be established.

The level of safety measures required depends on the magnitude of the risk posed by an AI system, which in turn depends on the system’s capabilities. Where it can be foreseen that the development process or the system itself will pose particularly high risks, it is crucial for safety measures to be developed and tested proactively.

- **Accuracy**

Accuracy pertains to an AI system’s ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models.

An explicit and well-formed development and evaluation process can support, mitigate and correct unintended risks from inaccurate predictions. When occasional inaccurate predictions cannot be avoided, it is important that the system can indicate how likely these errors are. A high level of accuracy is especially crucial in situations where the AI system directly affects human lives.

- **Reliability and Reproducibility**

It is critical that the results of AI systems are reproducible, as well as reliable. A reliable AI system is one that works properly with a range of inputs and in a range of situations. This is needed to scrutinise an AI system and to prevent unintended harms.

Reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions. This enables scientists and policy makers to accurately describe what AI systems do. Replication files can facilitate the process of testing and reproducing behaviours.

3. Privacy and data governance

Means that AI systems are developed and used in compliance with existing privacy and data protection rules, while processing data that meets high standards in terms of quality and integrity.

○ Including respect for privacy, quality and integrity of data, and access to data

Closely linked to the principle of prevention of harm is privacy, a fundamental right particularly affected by AI systems. Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be deployed, its access protocols and the capability to process data in a manner that protects privacy.

- **Privacy and data protection**

AI systems must guarantee privacy and data protection throughout a system’s entire lifecycle. This includes the information initially provided by the user, as well as the information generated about the user over the course of their interaction with the system (e.g. outputs that the AI system generated for specific users or how users responded to particular recommendations).

Digital records of human behaviour may allow AI systems to infer not only individuals’ preferences, but also their sexual orientation, age, gender, religious or political views. To allow individuals to trust the data gathering process, it must be ensured that data collected about them will not be used to unlawfully or unfairly discriminate against them.

- **Quality and integrity of data**

The quality of the data sets used is paramount to the performance of AI systems. When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. This needs to be addressed prior to training with any given data set. In addition, the integrity of the data must be ensured. Feeding malicious data into an AI system may change its behaviour, particularly with self-learning systems.

Processes and data sets used must be tested and documented at each step such as planning, training, testing and deployment. This should also apply to AI systems that were not developed in-house but acquired elsewhere.

- **Access to data**

In any given organisation that handles individuals' data (whether someone is a user of the system or not), data protocols governing data access should be put in place. These protocols should outline who can access data and under which circumstances. Only duly qualified personnel with the competence and need to access individual's data should be allowed to do so.

4. Transparency

Means that AI systems are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their rights.

○ Including traceability, explainability and communication

This requirement is closely linked with the principle of explicability and encompasses transparency of elements relevant to an AI system: the data, the system and the business models.

- **Traceability**

The data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible standard to allow for traceability and an increase in transparency. This also applies to the decisions made by the AI system. This enables identification of the reasons why an AI-decision was erroneous which, in turn, could help prevent future mistakes. Traceability facilitates auditability as well as explainability.

- **Explainability**

Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability).

Whenever an AI system has a significant impact on people’s lives, it should be possible to demand a suitable explanation of the AI system’s decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency).

- **Communication**

AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system. This entails that AI systems must be identifiable as such. In addition, the option to decide against this interaction in favour of human interaction should be provided where needed to ensure compliance with fundamental rights.

Beyond this, the AI system’s capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy, as well as its limitations.

Transparency stands as a cornerstone in the ethical development of Language Models. It involves openness, clarity, and accessibility of information regarding how these models are constructed, trained, and utilized. Transparent practices in LLM development offer several key advantages:

- ➔ **Trust and Credibility**: Transparency fosters trust among users, stakeholders, and the public. When the inner workings of LMs are made visible, it helps users understand the capabilities and limitations of the models, thus establishing credibility in their usage.

- ➔ **Identification and Mitigation of Biases**: Transparent practices facilitate the identification of biases within datasets or algorithms. By making the development process transparent, it becomes easier to detect and address biases, thereby promoting fairness and inclusivity in the models.

5. Diversity, non-discrimination and fairness

Means that AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law.

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

In order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. Besides the consideration and involvement of all affected stakeholders throughout the process, this also entails ensuring equal access through inclusive design processes as well as equal treatment. This requirement is closely linked with the principle of fairness.

- **Avoidance of unfair bias**

Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models.

The continuation of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation. Harm can also result from the intentional exploitation of (consumer) biases or by engaging in unfair competition, such as the homogenisation of prices by means of collusion or a nontransparent market.

Identifiable and discriminatory bias should be removed in the collection phase where possible. The way in which AI systems are developed (e.g. algorithms' programming) may also suffer from unfair bias.

This could be counteracted by putting in place oversight processes to analyse and address the system's purpose, constraints, requirements and decisions in a clear and transparent manner. Moreover, hiring from diverse backgrounds, cultures and disciplines can ensure diversity of opinions and should be encouraged.

- **Accessibility and universal design**

Particularly in business-to-consumer domains, systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. Accessibility to this technology for persons with disabilities, which are present in all societal groups, is of particular importance.

AI systems should not have a one-size-fits-all approach and should consider Universal Design principles addressing the widest possible range of users, following relevant accessibility standards.

This will enable equitable access and active participation of all people in existing and emerging computer-mediated human activities and with regard to assistive technologies.

- **Stakeholder Participation**

In order to develop AI systems that are trustworthy, it is advisable to consult stakeholders who may directly or indirectly be affected by the system throughout its life cycle.

It is beneficial to solicit regular feedback even after deployment and set up longer term mechanisms for stakeholder participation, for example by ensuring workers information, consultation and participation throughout the whole process of implementing AI systems at organisations.

Strategies for Ensuring Fairness and Inclusivity in Language Models:

Addressing biases and promoting fairness in language models requires a concerted effort from developers, researchers, policymakers, and society as a whole.

Some strategies to ensure fairness include:

- **Diverse Representation**: Actively seeking diverse perspectives in dataset creation, model development, and testing phases.

- **Continuous Monitoring and Auditing**: Implementing regular audits and checks for biases, with a focus on fairness metrics and real-world impact.

- **Stakeholder Engagement**: Involving diverse stakeholders in the development and decision-making processes to understand and address concerns.

- In addition, transparency and accountability are key. Making the development processes of language models transparent and being accountable for their outputs can help identify and rectify biases.

Creating fair and unbiased language models is not just a technical challenge but a moral imperative. As these models increasingly shape our interactions and decisions, addressing biases becomes crucial to ensure equitable outcomes.

By understanding biases, acknowledging their impact, and actively working towards fairness and inclusivity, we can pave the way for a more just and equitable technological landscape.

6. Social and environmental well-being

Means that AI systems are developed and used in a sustainable and environmentally friendly manner as well as in a way to benefit all human beings, while monitoring and assessing the long-term impacts on the individual, society and democracy.

○ Including sustainability and environmental friendliness, social impact, society and democracy

In line with the principles of fairness and prevention of harm, the broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system's life cycle.

- **Sustainability and ecological responsibility** of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, such as for instance the

Sustainable Development Goals. Ideally, AI systems should be used to benefit all human beings, including future generations. Sustainable and environmentally friendly AI. AI systems promise to help tackling some of the most pressing societal concerns, yet it must be ensured that this occurs in the most environmentally friendly way possible.

The system's development, deployment and use process, as well as its entire supply chain, should be assessed in this regard, e.g. via a critical examination of the resource usage and energy consumption during training, opting for less harmful choices. Measures securing the environmental friendliness of AI systems' entire supply chain should be encouraged.

- **Social impact**

Ubiquitous exposure to social AI systems in all areas of our lives (be it in education, work, care or entertainment) may alter our conception of social agency, or impact our social relationships and attachment. While AI systems can be used to enhance social skills, they can equally contribute to their deterioration. This could also affect people's physical and mental wellbeing. The effects of these systems must therefore be carefully monitored and considered.

- **Society and Democracy**

Beyond assessing the impact of an AI system's development, deployment and use on individuals, this impact should also be assessed from a societal perspective, taking into account its effect on institutions, democracy and society at large. The use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including not only political decision-making but also electoral contexts.

7. Accountability

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

The requirement of accountability complements the above requirements, and is closely linked to the principle of fairness. It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.

- **Auditability**

Auditability entails the enablement of the assessment of algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available. Evaluation by internal and external auditors, and the availability of such evaluation reports, can contribute to the trustworthiness of the technology. In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited.

- **Minimisation and reporting of negative impacts**

Both the ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome, must be ensured. Identifying, assessing, documenting and minimising the potential negative impacts of AI systems is especially crucial for those (in)directly affected. Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI system.

The use of impact assessments (e.g. red teaming or forms of Algorithmic Impact Assessment) both prior to and during the development, deployment and use of AI systems can be helpful to minimise negative impact.

These assessments must be proportionate to the risk that the AI systems pose.

- **Trade-offs**

When implementing the above requirements, tensions may arise between them, which may lead to inevitable trade-offs. Such trade-offs should be addressed in a rational and methodological manner within the state of the art.

This entails that relevant interests and values implicated by the AI system should be identified and that, if conflict arises, trade-offs should be explicitly acknowledged and evaluated in terms of their risk to ethical principles, including fundamental rights. In situations in which no ethically acceptable trade-offs can be identified, the development, deployment and use of the AI system should not proceed in that form.

Any decision about which trade-off to make should be reasoned and properly documented. The decision-maker must be accountable for the manner in which the appropriate trade-off is being made, and should continually review the appropriateness of the resulting decision to ensure that necessary changes can be made to the system where needed.

- **Redress**

When unjust adverse impact occurs, accessible mechanisms should be foreseen that ensure adequate redress.

Knowing that redress is possible when things go wrong is key to ensure trust. Particular attention should be paid to vulnerable persons or groups.

The decision-making process involved in training datasets and algorithms significantly impacts the performance and ethical implications of LMs. It's imperative to illuminate how these decisions are made:

- Dataset Selection and Curation: Transparency in dataset selection involves disclosing the sources, composition, and potential biases within the data used to train the models. Furthermore, explaining the criteria for inclusion or exclusion of certain data sets ensures a clearer understanding of the model's learning process.

- Algorithmic Choices and Training Methods: Describing the algorithms employed and the rationale behind their selection helps users comprehend the model's behavior. This includes detailing how the models learn, adapt, and make predictions based on the provided data.

Accountability measures are crucial to ensure responsible and ethical use of Language Models. This involves establishing mechanisms to monitor, assess, and rectify potential issues arising from their actions and outputs:

- Ethical Guidelines and Governance: Formulating and adhering to ethical guidelines and governance frameworks sets the stage for accountable LMs. This includes defining responsible use, outlining consequences for misuse, and establishing oversight bodies to enforce compliance.

- Continuous Evaluation and Auditing: Regular evaluation and auditing of LMs are essential to identify biases, errors, or unintended consequences. This process enables prompt corrective actions and improvements to maintain ethical standards.

- Transparency Reports: Publishing transparency reports that document the model's performance metrics, data sources, and any incidents of misuse contributes to accountability. This transparency aids in building trust and allows for external scrutiny.

Useful links

Checklist (ALTAI):

https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment