The EU "Framework for Trustworthy AI"

Roberto V. Zicari

Z-Inspection® Initiative

http://z-inspection.org

SNU, September 20 and September 27, 2022

Lessons Content

1. **Why Trustworthy AI?**

2. The EU "Framework for Trustworthy AI"

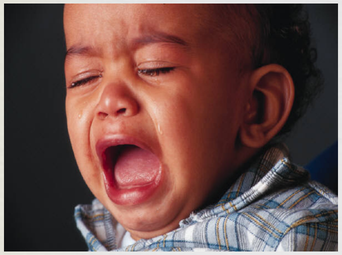
3. EU AI Act proposed law
(*Guest Lecture* on November 8)

.

1. Why Trustworthy AI?

“Everything we love about civilization is a product of intelligence, so amplifying our human intelligence with artificial intelligence has the potential of helping civilization flourish like never before **– as long as we manage to keep the technology *beneficial.*“**

-- Max Tegmark, President of the Future of Life Institute



Do no harm
Can we explain decisions?

What if the decision made using AI-driven algorithm harmed somebody, and you cannot explain how the decision was made?

How do we know when AI is “beneficial”?

- **Appropriate use:** Assess if the data and algorithm are appropriate to use for the purpose anticipated and perception of use.
  - Suppose we assess that the AI is technically *unbiased* and *fair* –this does not imply that it is acceptable to deploy it.
- **Remedies**: If **risks** are identified, define ways to mitigate risks (when possible)
- **Ability to redress**

2. EU Framework for Trustworthy Artificial Intelligence

- „**Trust** in the development, deployment and use of AI systems concerns **not only the technology's** inherent properties, but also the qualities of the **socio-technical systems** involving AI applications."

Source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

EU Framework for Trustworthy Artificial Intelligence

EU High-Level Expert Group on AI defined **ethics *guidelines* for *trustworthy* artificial intelligence:**

(1) **lawful** - respecting all applicable laws and regulations

(2) **ethical** - respecting ethical principles and values

(3) **robust** - both from a technical perspective while taking into account its social environment

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Fundamental rights as moral and legal entitlements

The EU “believes in an approach to AI ethics based on the **fundamental rights** enshrined in
the EU Treaties
the EU Charter and
international human rights law.

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

EU treaties

- "The EU treaties are binding agreements between EU member countries. They set out EU objectives, rules for EU institutions, how decisions are made and the relationship between the EU and its member countries. Every action taken by the EU is founded on treaties."

- Source: https://eur-lex.europa.eu/collection/eu-law/treaties/treaties-force.html

The EU Charter of Fundamental Rights,

- “The Charter of Fundamental Rights of the European Union (the Charter) brings together **the fundamental rights of everyone living in the European Union (EU).** It was introduced to bring consistency and clarity to the rights established at different times and in different ways in individual EU Member States.”
- Source: https://www.equalityhumanrights.com/en/what-are-human-rights/how-are-your-rights-protected/what-charter-fundamental-rights-european-union

European Convention on Human Rights

- The **European Convention on Human Rights** is an international convention to protect human rights and political freedoms in Europe.
- The Convention established the European Court of Human Rights.
- Any person who feels their rights have been violated under the Convention by a state party can take a case to the Court. Judgments finding violations are binding on the States concerned and they are obliged to execute them.

Source : Wikipedia https://en.wikipedia.org/wiki/European_Convention_on_Human_Rights

United Nations International human rights law

- "International human rights law lays down obligations which States are bound to respect. By becoming parties to international treaties**, States assume obligations and duties under international law to respect, to protect and to fulfill human rights.** The obligation to respect means that States must refrain from interfering with or curtailing the enjoyment of human rights. The obligation to protect requires States to protect individuals and groups against human rights abuses. **The obligation to fulfill means that States must take positive action to facilitate the enjoyment of basic human rights."**

- **Source:https://www.ohchr.org/en/instruments-and-mechanisms/international-human-rights-law**

Fundamental rights as moral and legal entitlements

“ **Respect for fundamental rights,** within a framework **of democracy and the rule of law**, provides the most promising foundations **for identifying abstract ethical principles and values,** which can be **operationalised in the context of AI.** “

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Fundamental rights as moral and legal entitlements (cont.)

- "The EU Treaties and the EU Charter prescribe a series of fundamental rights that **EU member states and EU institutions are legally obliged to respect when implementing EU law.** These rights are described in the EU Charter by reference **to dignity, freedoms, equality and solidarity, citizens' rights and justice.**

- The common foundation that unites these rights can be understood as rooted in respect for **human dignity** – thereby reflecting what we describe as a **"human-centric approach"** in which the human being enjoys a unique and inalienable moral status of primacy in the civil, political, economic and social fields."

Fundamental rights as moral and legal entitlements (cont.)

- „While the rights set out in the EU Charter are legally binding, it is important to recognise that fundamental rights do not provide comprehensive legal protection in every case.
- **For the EU Charter, for instance, it is important to underline that its field of application is limited to areas of EU law.**
- **International human rights law** and in particular the **European Convention on Human Rights** are legally binding on EU Member States, including in areas that fall outside the scope of EU law. „

Fundamental rights as moral and legal entitlements (cont.)

- „At the same time, **fundamental rights are also bestowed on individuals and (to a certain degree) groups by virtue of their moral status as human beings, independently of their legal force.**

- Understood as legally enforceable rights, **fundamental rights therefore fall under the first component of Trustworthy AI (lawful AI)**, which safeguards compliance with the law. “

Fundamental rights as moral and legal entitlements (cont.)

"Understood as the rights of everyone, rooted in the inherent moral status of human beings, they also underpin the second component of Trustworthy AI (**ethical AI**), dealing with **ethical norms that are not necessarily legally binding** yet crucial to ensure trustworthiness".

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

“Lawful” AI is not dealt by the EU guidelines.

- The EU document “**does not aim to offer guidance on the legal componen**t, for the purpose of these non-binding guidelines, references to fundamental rights reflect the latter component. „

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

From fundamental rights to ethical principles

Four ethical principles, rooted in fundamental rights

(i) Respect for human autonomy
(ii) Prevention of harm
(iii) Fairness
(iv) Explicability

- There may be **tensions** between these principles.

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

(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.** "

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019

The principle of respect for human autonomy (cont.)

- “ 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**.

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019

The principle of respect for human autonomy (cont.)

“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

8 April, 2019.

- "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 secur**e. They must be **technically robust** and it should be ensured that they are not open to malicious use."

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019

The principle of prevention of harm (cont.)

- “ **Vulnerable persons should receive greater attention** and be included in the development, deployment and use of AI systems.”

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019

The principle of prevention of harm (cont.)

- “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

April, 2019.

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

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8

Different interpretations of fairness

There are many different interpretations of fairness;

Fairness has both a

(1) **substantive**
and a
(2) **procedural** dimension.

The principle of fairness (cont).

(1) The **substantive** dimension implies a commitment to:

“Equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation.

The principle of fairness (cont)

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."

The principle of fairness (cont.)

- **“The use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice. “**

The principle of fairness (cont.)

- “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.”

Proportionality

- “**Proportionality** is a core principle in **international law, which provides that the legality of an action shall be determined depending on the respect of the balance between the objective and the means and methods used as well as the consequences of the action**. This principle implies an obligation **to appreciate the context** before deciding on the legality or the illegality of an action.
- This assessment is the responsibility of those who act. In case of dispute or doubt, tribunals can assess the facts and their legality a posteriori.”

Source: https://guide-humanitarian-law.org/content/article/3/proportionality/

The principle of fairness (cont.)

(2) 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**.

The principle of fairness (cont.)

Fairness is closely linked to the rights to:

Non-discrimination, Solidarity and Justice

Non-discrimination

- **Definition of non-discrimination:**
  the practice of treating people, companies, countries, etc. in the same way as others in order to be **fair**.

Source: Merriam-Webster

Solidarity

- **Definition of *solidarity* :**

unity (as of a group or class) that produces or is based on community of interests, objectives, and standards

Source: Merriam-Webster

Justice

- **Definition of *justice*:**

the maintenance or administration of what is just especially by the impartial adjustment of conflicting **claims** or the assignment of merited rewards or punishments

Source: Merriam-Webster

The four pillars of medical ethics

- **Beneficence (doing good)**
- **Non-maleficence (to do no harm)**
- **Autonomy (giving the patient the freedom to choose freely, where they are able)**
- **Justice (ensuring fairness)**

(iv)The principle of explicability

This is a new Ethical principle…

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

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, April, 2019

The principle of explicability (cont.)

- 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. “

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, April, 2019

The principle of explicability (cont.)

- **Explicability and Responsibility** are closely linked to the rights relating to **Justice.**

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, April, 2019

The principle of explicability (cont.)

- “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**. “

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, April, 2019

The principle of explicability (cont.)

- “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. “

Tensions and Trade-offs

- „**Tension**s may arise between the above ethical principles**, for which there is no fixed solution**.

- In line with the EU fundamental commitment to democratic engagement, due process and open political participation, ***methods of accountable deliberation to deal with such tensions should be established.*** “

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019

Example of Tensions and Trade-offs

- “In various application domains, *the principle of prevention of harm* and *the principle of human autonomy* may be in conflict.

- Consider as an example the use of AI systems for ‘predictive policing’, which may help to reduce crime, but in ways that entail surveillance activities that impinge on individual liberty and privacy. “

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019

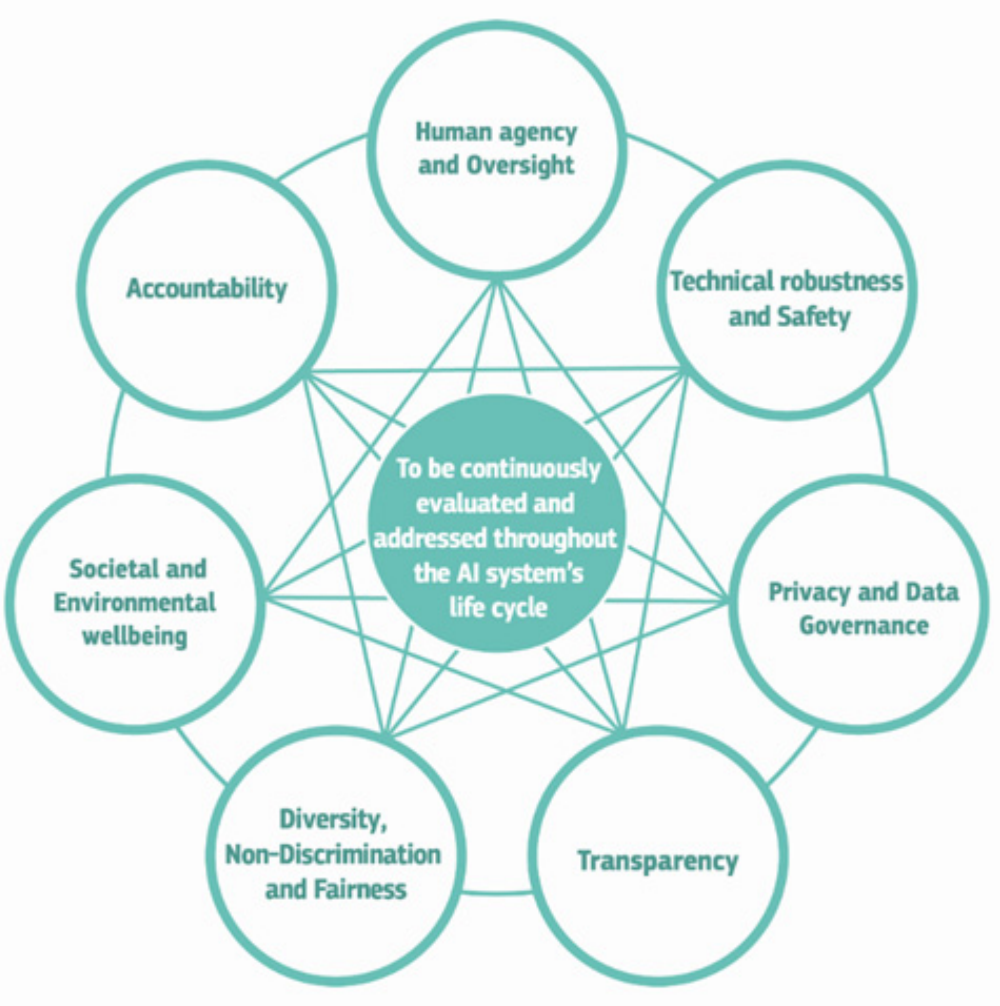
Tensions and Trade-offs (cont.)

- **“ AI practitioners can hence not be expected to find the right solution based on the ethical principles, yet they should approach ethical dilemmas and trade-offs via reasoned, evidence-based reflection rather than intuition or random discretion.**

- **There may be situations, however, where no ethically acceptable trade-offs can be identified.“**

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019

7 Requirements of Trustworthy AI


Human agency and Oversight
Technical robustness and Safety
Privacy and Data Governance
Transparency
Diversity, Non-Discrimination and Fairness
Societal and Environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle

AI System Life Cycle

- Design
- Development
- Deployment
- Monitoring
- Termination

Seven Requirements (+ sub-requirements) for Trustworthy AI

1 Human agency and oversight

***Including fundamental rights**, human agency and human oversight*

2 Technical robustness and safety

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

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Requirements for Trustworthy AI

3 Privacy and data governance

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

4 Transparency

Including traceability, explainability and communication

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Requirements of Trustworthy AI

5 Diversity, non-discrimination and fairness

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

6 Societal and environmental wellbeing

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

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Requirements of Trustworthy AI

7 Accountability

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

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

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;
- **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. “

1. Human agency and oversight

- **All potential impacts that AI systems may have on** ***fundamental rights*** **should be accounted for** and that ***the human role in the decision- making process is protected.***

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021..

Human agency and oversight (cont.)

- “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.*”

Human agency and oversight (cont.)

- ***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”

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Fundamental rights (cont.).

- "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. "

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Fundamental rights (cont.)

- “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.”

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Human agency and oversight (cont.)

- ***Human agency.*** ***(Role)***

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.

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Human agency (cont.)

“ 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**. “

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Human agency (cont.)

“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.”

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Human agency and oversight (cont.)

- ***Human oversight*.** (Control)

"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-the- loop (HITL), **human-on-the-loop** (HOTL), **or human-in-command** (HIC) approach. "

Human oversight (cont.)

- **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.

Human oversight (cont.)

„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. "

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Human oversight (cont.)

“ 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. „

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

2. *Technical robustness and safety*

- AI systems should be secure and resilient in their operation in a way that *minimizes potential harm*, optimizes accuracy, and fosters confidence in their reliability;

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021..

Technical robustness and safety (cont.)

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

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Technical robustness and safety (cont.

- "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. "

Technical robustness and safety (cont.)

- ***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. "

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Resilience to attack and security. (cont.)

- “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. “**

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Resilience to attack and security. (cont.)

“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.”

Technical robustness and safety (cont.)

***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**"

Fallback plan and general safety. (cont.)

“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. “

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Technical robustness and safety (cont.)

- *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**. “

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Accuracy (cont.)

- “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. “**

Technical robustness and safety (cont.)

- ***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.** "

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

.

Technical robustness and safety (cont.)

- **“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

- Given the vast quantities of data processed by AI systems, this principle impresses *the importance of protecting the privacy, integrity, and quality of the data and protects human rights of access to it;*

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications.

Privacy and data governance (cont.)

- “Closely linked to the *principle of prevention of harm* is **privacy**, a fundamental right particularly affected by AI systems. “

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Privacy and data governance (cont.)

- “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 protocol**s and the capability to process data in a manner that protects privacy. “

Privacy and data governance (cont.)

- ***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). "

Privacy and data governance (cont.)

“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. “

Privacy and data governance (cont.)

- ***Quality and integrity of data*.**

"The **quality of the data se**ts used is paramount to the performance of AI systems. When data is gathered, it may contain **socially constructed biases, inaccuracies, errors and mistakes."**

source: *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Privacy and data governance (cont.)

“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. “

Privacy and data governance (cont.)

- ***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

- AI systems need to be *understandable* at a human level so that decisions made through AI can be traced back to their underlying data. *If a decision cannot be explained it cannot easily be justified;*

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021.

Transparency (cont.)

- “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. “**

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Transparency (cont.)

- ***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 **auditabilit**y as well as explainability. “

Transparency (cont.)

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. “

Transparency (cont.)

- **“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)”**

- **source:** *Ethics Guidelines for Trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence. European commission, 8 April, 2019.

Explainability (cont.)

“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). “

Transparency (cont.)

- ***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. “

Communication (cont.)

- “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. “**

5. Diversity, non-discrimination, and fairness

- *AI systems need to be inclusive and* non-biased in their application. This is challenging when the data is not reflective of all the potential stakeholders of an AI system;

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications

Diversity, non-discrimination, and fairness (cont.)

- "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.***

Diversity, non-discrimination, and fairness (cont.)

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."

Avoidance of unfair bias. (cont.)

- **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 non-transparent 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.

Diversity, non-discrimination, and fairness (cont.)

- ***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."

Diversity, non-discrimination, and fairness (cont.)

- *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. "

6. Societal and environmental wellbeing

- In acknowledging the potential power of AI systems, this principle emphasizes the need for *wider social concerns*, including the environment, democracy, and individuals to be taken into account;

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021. Manuscript submitted for publications

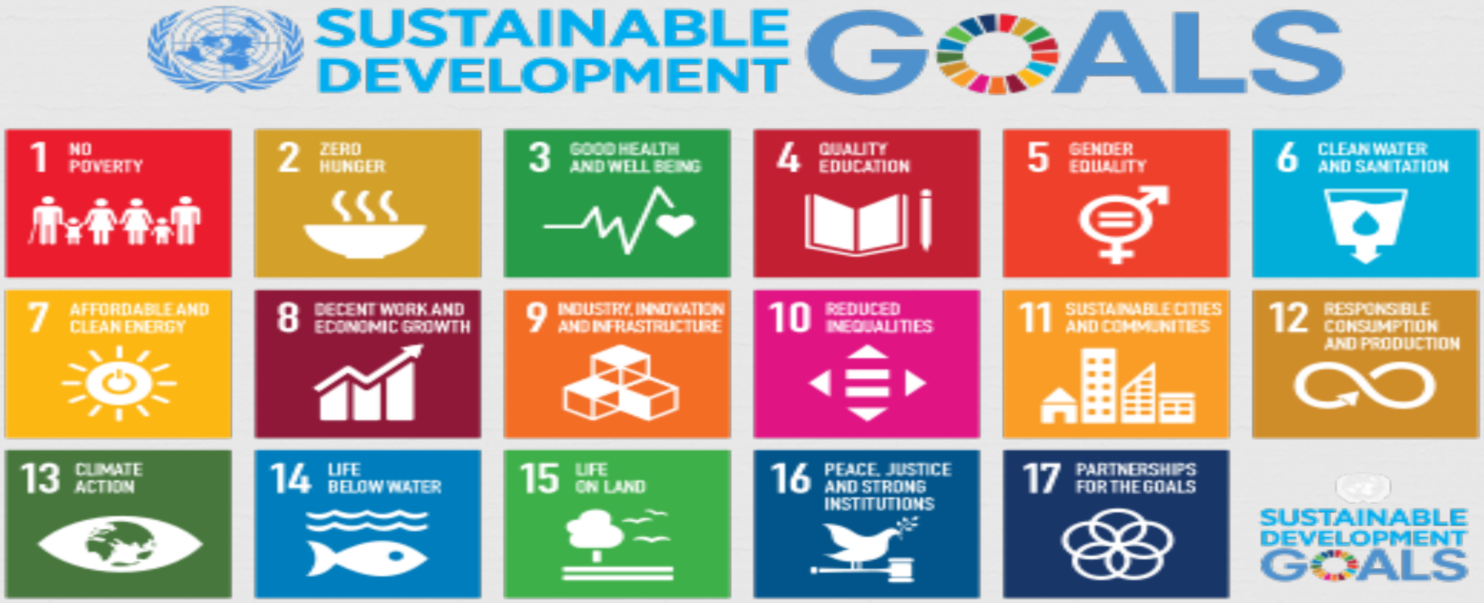
Societal and environmental wellbeing (cont.)

- “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.”**

Societal and environmental wellbeing (cont.)

- “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.”

UN Sustainable Development Goals


SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

17 Sustainable Development Goals

- GOAL 1: No Poverty
- GOAL 2: Zero Hunger
- GOAL 3: Good Health and Well-being
- GOAL 4: Quality Education
- GOAL 5: Gender Equality
- GOAL 6: Clean Water and Sanitation
- GOAL 7: Affordable and Clean Energy
- GOAL 8: Decent Work and Economic Growth
- GOAL 9: Industry, Innovation and Infrastructure
- GOAL 10: Reduced Inequality
- GOAL 11: Sustainable Cities and Communities
- GOAL 12: Responsible Consumption and Production
- GOAL 13: Climate Action
- GOAL 14: Life Below Water
- GOAL 15: Life on Land
- GOAL 16: Peace and Justice Strong Institutions
- GOAL 17: Partnerships to achieve the Goal

Societal and environmental wellbeing (cont.)

- ***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. "

Societal and environmental wellbeing (cont.)

- ***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. “**

Societal and environmental wellbeing (cont.)

- ***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

- This principle, rooted **in *fairn*es**s, seeks to ensure ***clear lines of responsibility and accountability*** for the outcomes of AI systems, mechanisms for addressing trade-offs, and an environment in which concerns can be raised.

- However, the interpretation, relevance, and implementation of trustworthy AI *depends on the domain and the context* where the AI system is used.

- Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021.

7. Accountability

- “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 accountab**ility for AI systems and their outcomes, **both before and after their development, deployment and use**. “

Accountability (cont.)

- ***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. "

Accountability (cont.)

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.

- ***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. “

Trade-offs (cont.)

- “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. “**

Trade-offs (cont.)

- "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."

Accountability (cont.)

- ***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. “

Challenges and Limitations

The AI HLEG trustworthy AI *guidelines are not a law and are not contextualized by the domain they are involved in.* The meaning of some of the seven requirements is not anchored to the context (e.g., fairness, wellbeing, etc.).

They mainly offer a static checklist (AI HLEG 2020) and do not take into account changes of the AI over time.

Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021

Challenges and Limitations

They do not distinguish different applicability of the AI HLEG trustworthy AI guidelines (e.g., during design vs. after production) as well as different stages of algorithmic development, starting from business and use-case development, design phase, training data procurement, building, testing, deployment, monitoring.

There are not available best practices to show how to implement such requirements and apply them in practice.

The guidelines do not explicitly address the lawful part of the assessment.

Source: On Assessing Trustworthy AI in Healthcare . Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. Roberto V. Zicari, et al 2021.