Move theLF-AI Trusted AI - Principles - Draft

**Some important links**

* LF-AI Committee link <https://wiki.lfai.foundation/>
* LF-AI Trusted AI Committee [https://wiki.lfai.foundation/display/DL/Trusted+AI+Committee](https://wiki.lfai.foundation/display/DL/Trusted%2BAI%2BCommittee)
* LF-AI

Principles for trusted AI have been defined by many communities and entities, and they can be represented as a Venn diagram because they overlap significantly in many ways. The following principles are intended to be considered when projects are evaluated by the LF-AI to check their robustness with respect to these dimensions. They are high level principles that act as guides.

We have taken the definitions from these documents:

* [EU] Ethics Guidelines for Trustworthy AI - High-Level  Expert Group on Artificial Intelligence set up by the European Commission <https://ec.europa.eu/futurium/en/ai-alliance-consultation>
* [EUFeb] On Artificial Intelligence -A European approach to excellence and tru [https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020\_en.pdf](https://urldefense.proofpoint.com/v2/url?u=https-3A__ec.europa.eu_info_sites_info_files_commission-2Dwhite-2Dpaper-2Dartificial-2Dintelligence-2Dfeb2020-5Fen.pdf&d=DwMGaQ&c=jf_iaSHvJObTbx-siA1ZOg&r=PPBt9g5r27h23QHszTu1gNAeqCepqpCmqLUK9BT6bNk&m=VE_fZQ7vYd_woY_F_c5hMNMl0ACWPS7Lmfx8Wb7tUBQ&s=37EYkmcv_FZULjtd72Ojj-RAaasDn-kctehlPOCw4VM&e=)
* [IEEE] Ethically Aligned Design, IEEE <https://ethicsinaction.ieee.org/>
* [DoD] AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense <https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF>

The following are the principles:

 Equitability

  The provider or designer should take deliberate steps to avoid unintended bias, acting with fairness, in the development and deployment of AI systems that would inadvertently cause harm to persons.

    *Related terms :  Positive Impact on Society****,*** *Negative Bias, Fair, Human Rights*

*Definition taken and adapted from :  [DOD]*

Reproducibility

    AI engineering discipline should be sufficiently advanced such that technical experts possess an appropriate understanding of the technology, development processes, and operational methods of its AI systems, including transparent and auditable methodologies, data sources, and design procedure and documentation.

     *Related terms: Performant, Available, Transparent, Traceable*

*Definition taken from :  [DOD]*

 Transparency

Transparency relates to information regarding the presence of AI and Algorithmic transparency where relevant and especially in high-risk AI systems. Ensuring clear information is provided about AI system’s capabilities and limitations, in particular the purpose for which the systems are intended, training and testing data sets where feasible, the conditions under which they can be expected to function as intended and the expected level of accuracy in achieving the specified purpose. In addition, where it is not immediately obvious, users should be clearly informed when they are interacting with an AI system and not a human being.

 *Related terms: Explainability*

 *Definition derived from: [EUFeb]*

Robustness

Robustness: refers to stability, resilience and performance of the systems and machines dealing with moving ecosystems. To be more defined in the context of AI.

     *Related terms: Reliable, Stable, Performant, Resilient*

*Definition taken from :  [EU]*

Governance

AI systems should be designed and engineered to fulfill their intended function while possessing the ability to detect and avoid unintended harm or disruption, and for human or automated disengagement or deactivation of deployed systems that demonstrate unintended escalatory or other behavior.
     *Related terms: Explainable*

*Definition taken from :  [DOD]*

Privacy

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.

         *Related terms: Data Security, Data Agency*

*Definition taken from :  [EU]*

Security

AI systems should have an explicit, well-defined domain of use, and the safety, security, and robustness of such systems should be tested and assured across their entire life cycle within that domain of use.

 *Related terms : Responsible, Safe, Traceable, Robust*

         *Definition taken from :  [DOD]*

Accountability

Human beings should exercise appropriate levels of judgment and remain responsible for the development, deployment, use, and outcomes of AI systems.

        *Related terms : Responsible*

         *Definition taken from :  [DOD]*

----------------------------------------------------------------------------------------------------------------------

Table mapping the terms to the submissions from the LF-AI Trusted AI Group

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term**  | **AT&T** | **IBM** | **Orange** | **Tencent** |
| **Applies to all Principles** |  |  | Orange1. Foundation based on principles  | Tencent - The “Ethical Ark” (ARCC) for the healthy and safe development of AI |
|  **Equitable** | ATT1. Should be by people for people | IBM1. Fair |  |  |
| **Reproducible** |  |  | Orange3. Evaluation and organization of an IA life cycle management (using checklist methodology) | Tencent1. AvailableTencent4. Controllable |
| **Governable** |  | IBM2. Explainable | Orange2  Realisation using technical and governance tools     | Tencent3. ComprehensibleTencentTrust1. Laws and regulationsTencentTrust2. Industry self regulation |
| **Private** | ATT2. Accessible and shared |  |  |  |
| **Secure** | ATT3. Secure and ethical | IBM3. Robust |  | Tencent2. Reliable |
| **Accountable** | ATT3. Secure and ethical | IBM4. Transparent and Accountable |  |  |
|  |  |  |  | TencentTrust3. Education and awareness |

**References and Related Materials**

 wikipedia <https://en.wikipedia.org/wiki/Artificial_intelligence>

LF-AI Link <https://wiki.lfai.foundation/>

Human Compatible <https://techcrunch.com/2019/10/06/an-interview-with-dr-stuart-russell-author-of-human-compatible-artificial-intelligence-and-the-problem-of-control/>

 European AI High Level Expert Group : ethics guidelines for trustworthy AI & policy and investment recommendations for trustworthy AI<https://lists.lfai.foundation/g/trustedai-committee/files/EU-Ethics-Guidelines-for-Trustworthy-AI.pdf>

**The AI-development Connection - A View from the South**

https://dl.acm.org/doi/pdf/10.1145/3375627.3377139?accessTab=true

--------------------------------------------------------------------------------------------------------------------------

**Principles extracted from the  member submissions**

From AT&T  - AI Guiding Principles

o   ATT1. Should be by people for people

o   ATT2. Accessible and shared

o   ATT3. Secure and ethical

From the EU-Ethics-Guidelines-for-Trustworthy-AI Trustworthy AI has three components, which should be met throughout the system's entire life cycle:

·         EU1. it should be lawful, complying with all applicable laws and regulations;

·         EU2. it should be ethical, ensuring adherence to ethical principles and values; and

·         EU3. it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm

From IBM – AI should be:

o   IBM1. Fair

o   IBM2. Explainable

o   IBM3. Robust

o   IBM4. Transparent and Accountable

From Orange – there are three steps to trustworthy AI:

-          Orange1. Foundation based on principles

-          Orange2  Realisation using technical and governance tools

-      Orange3. Evaluation and organization of an IA life cycle management (using checklist methodology)

From Orange – there is a requirement for an oversight agency that ensures that AI is:

OrangeOversight1.  Lawful - respecting all applicable laws and regulations

OrangeOversight2.  Ethical - respecting ethical principles and values

OrangeOversight3   Robust - both from a technical perspective while taking into account social environment

From Tencent - The “Ethical Ark” (ARCC) for the healthy and safe development of AI:

o   Tencent1. Available

o   Tencent2. Reliable

o   Tencent3. Comprehensible

o   Tencent4. Controllable

From Tencent – To build trust in AI need the following

o   TencentTrust1. Laws and regulations

o   TencentTrust2. Industry self regulation

o   TencentTrust3. Education and awareness