Linux Foundation AI Principles

LF-AI and Data Foundation - Principles Working Group, Trusted AI Committee
December 2, 2020

LF AI & DATA

Agenda

The Team

The Principles

- Why are the Principles important?
- The Principles Summary
- Methodology or Process
- More on Principles
- Next Steps
- References

-

Principles Working Group Team, Trusted AI Committee

Principles Working Group Team:

- Souad Ouali (Orange)
- Jeff Cao (Tencent)
- Francois Jezequel (Orange)
- Sarah Luger (Orange)
- Susan Malaika (IBM)
- Alka Roy (The Responsible Innovation Project/ex-AT&T)
- Alejandro Saucedo (The Institute for Ethical AI)
- Marta Ziosi (AI for People)

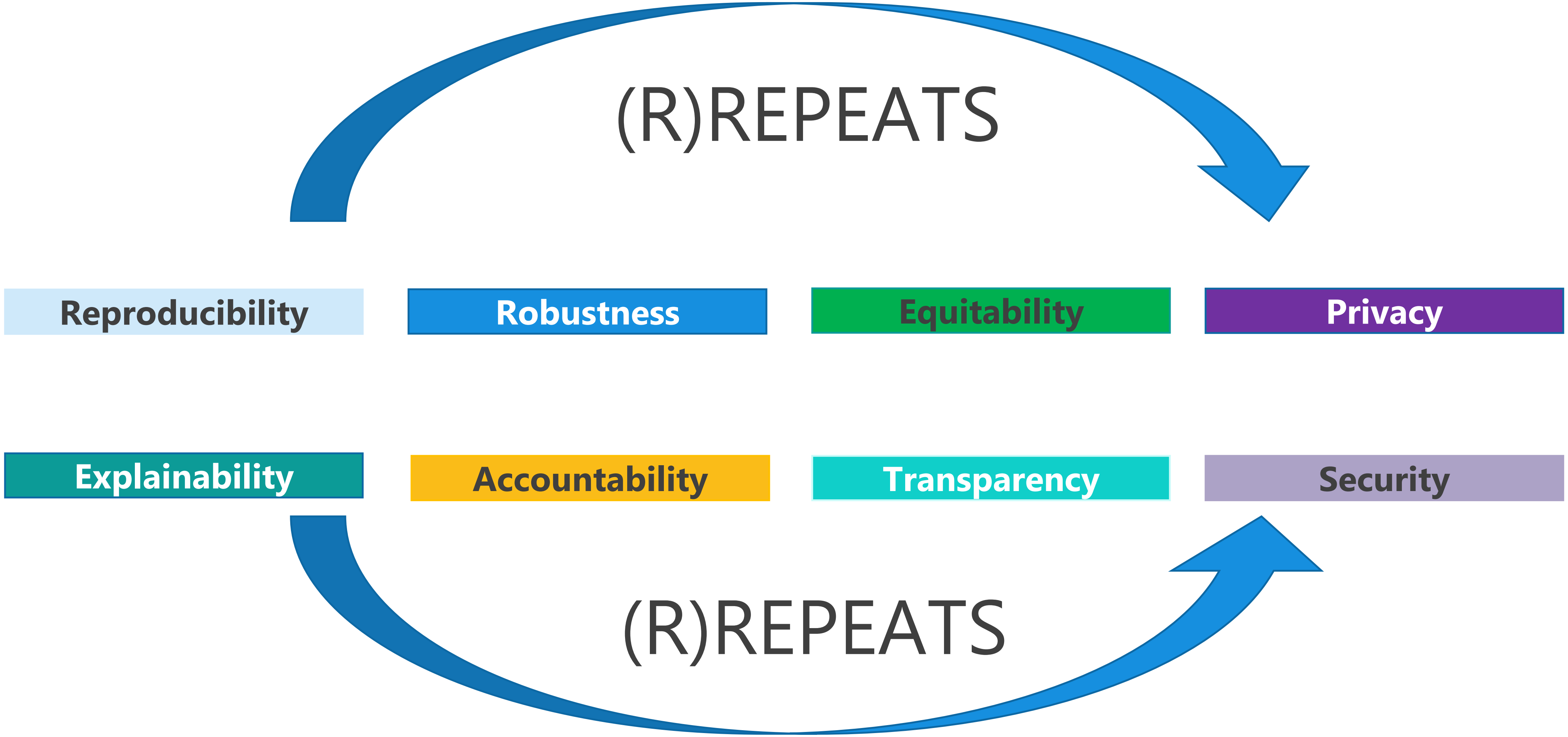
Why are the Principles important ?

- They encourage **TRUST** in the **DEVELOPMENT** of AI

- They can be **UNIVERSALLY SHARED** and **APPLIED** across regions, cultures and moral values

- They are **SIMPLE** and **EASY** to understand, and can be implemented in projects with flexibility to help ensure their adoption

The Principles – (R)REPEATS!



(R)REPEATS

Reproducibility

Robustness

Equitability

Privacy

Explainability

Accountability

Transparency

Security

(R)REPEATS

Methodology or Process

- Referenced influential and existing AI Principles & guidelines [listed on slide 19]
- Reviewed partner company's AI Principles and mapped across the industry / non-profit guidelines.
- Flat peer review structure across diverse set of Principles group members
- Key criteria for inclusion: Consensus across competing interests balanced with the need for open and innovative technology build with trust and accountability.

Reproducibility

- ✓ Reproducibility: Ability of an independent team to replicate in an equivalent AI environment, domain or area, the same experiences or results using the same AI methods, data, software, codes, algorithms, models, and documentation, to reach the same conclusions as the original research or activity. Adhering to this principle will ensure the reliability of the results or experiences produced by the AI.

- ✓ Beyond the definition :
  - ✓ We need to reach the same conclusions (not necessarily the same results) to ensure that AI is reliable and can be trusted.
  - ✓ Reproducibility is essentially connected to transparency, trust and scientific research. To ensure the quality of research, reproducibility is essential.

Reproducibility

- ✓ Beyond the definition :

  - ✓ The notion comes from this idea that any scientific experience and its results should be reproducible to be acknowledged. To ensure trust in AI and an AI fairness development, AI should be able to show that any team using the same tools and methods should be able to achieve the same results. It will give weight to the methods used in AI and ensure trust.

  - ✓ An AI experience whose results are reproducible under the same conditions, means that the data and methods used, were appropriate and may be reused for another AI experience, ensuring trust in the tools, methods and ways of collecting the data. It doesn't mean that the results will necessarily be correct (mistakes even with the best methods and data occur), it just guarantees that AI's development was made using rigorous and proper methods and data.

Reproducibility

- ✓ Beyond the definition :
  - ✓ Reproducibility will ensure reliability of the methods and the data used. And that anyone - under the same conditions of course - can obtain the same results. It will reinforce trust in AI, especially as many machine learning systems are black boxes even to the researchers that build them. That makes it hard for others to assess the results.
  - ✓ Being able to reproduce the way AI works, will reinforce the idea that these machines are predictable and under humans' control. Reproducibility will help give a moral guarantee. This principle is also deeply connected to explainability. Explainability is much more about "how" whereas reproducibility is more about "what was used", demonstrating that scientific methods are being used to ensure correct results.

Robustness

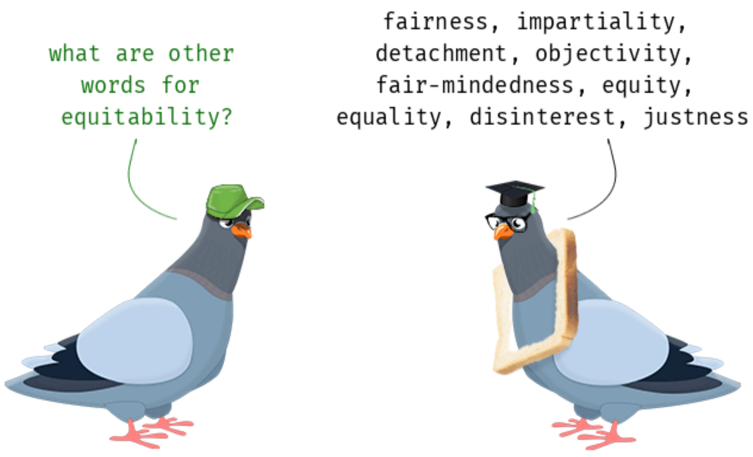
✓ Robustness: refers to stability, resilience and performance of the systems and machines dealing with changing ecosystems. AI must function in a robust way throughout their life cycles and potential risks should be continually assessed and managed.

- Beyond the definition:
  - Robustness: refers to stability, resilience to disruption and reliability in performance of the systems and machines dealing with a changing environment. AI should be designed to perform in a secure manner with meaningful safeguards to prevent the alteration of the system through either purposeful tampering in service or alteration of conditions away from the original assumptions around which the system was designed.
  - A robust AI is more trustable (capable of being trusted). Because a robust AI is stable, resilient and reliable during its life cycle it would be difficult to change its parameters.
  - This principle is connected to security.

Equitability

- Equitability: the AI and the people behind the AI should take deliberate steps - in the AI life-cycle - to avoid intended or unintended bias and unfairness that would inadvertently cause harm.

- Beyond the definition :


what are other words for equitability?
fairness, impartiality, detachment, objectivity, fair-mindedness, equity, equality, disinterest, justness

Thesaurus.plus

- The idea behind this principle is that biases are part of the human-kind and can't be avoided. Life is full of biases and we need to be aware of that. Sometimes biases are also useful for the AI progress. What is important is to be aware of that and put in place systems that will avoid harming people even inadvertently. It means being able to identify a bias harming people or which may harm people to prevent it. Equitability has to take place during the whole AI life cycle.

Privacy

- The AI must guarantee privacy and data protection throughout a system’s entire lifecycle. The lifecycle activities include the information initially collected from users, as well as information generated about users over the course of their interaction with the system e.g., outputs that the AI generated for specific users or how users responded to recommendations.
  The AI must ensure that data collected or inferred about individuals will not be used to unlawfully or unfairly discriminate against them. Privacy and transparency are especially needed when dealing with digital records that allow inferences such as identity, preferences, and future behavior.

- Beyond the definition :

  - Privacy means trust. It is connected to security. If the public knows that the AI protects its private data and uses it only for the purpose described by the AI, they will have more trust in AI, ensuring its progress and development to the benefit of humankind.

Explainability

- ✓ Explainability is the ability to describe how AI works, i.e., makes decisions. Explanations should be produced regarding both the procedures followed by the AI (i.e., its inputs, methods, models, and outputs) and the specific decisions that are made. These explanations should be accessible to people with varying degrees of expertise and capabilities including the public. For the explainability principle to take effect, the 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 the ability to explain the sources and triggers for decisions through transparent, traceable processes and auditable methodologies, data sources, and design procedure and documentation.

- ✓ Beyond the definition :

  - ✓ This principle is connected to transparency. Transparency is the idea to easily understand how AI works, and for what purposes. Explainability is understanding why AI makes a decision.

›

Accountability

- Accountability: requires the AI and people behind the AI to explain, justify, and take responsibility for any decision and action made by the AI. Mechanisms, such as governance and tools, are necessary to achieve accountability.

- Beyond the definition:

- This is the idea of responsibility, being responsible for what was done so AI, and the people behind the AI, can answer for it, explain and in the end control it.

Transparency

- ✓ Transparency: entails the disclosure around AI systems to ensure that people understand AI-based outcomes especially in high-risk AI domains. When relevant and not immediately obvious, users should be clearly informed when and how they are interacting with an AI and not a human being. For transparency, ensuring that clear information is provided about the AI’s capabilities and limitations, in particular the purpose for which the systems are intended, is necessary. Information about training and testing data sets where feasible, the conditions under which the AI can be expected to function as intended and the expected level of accuracy in achieving the specified purpose, should also be supplied.

- ✓ Beyond the definition:

  - ✓ AI should be easily understandable by people with varying degrees of expertise. Something we know is less feared and creates less frictions. Ignorance is the fertile soil for conflicts. If we know how AI is made, how it works and why decision are taken, we better understand AI. With transparency it is easier to respond to opposition to AI by focusing on how to change and enhance AI. If we ensure transparency, we work for trust in AI and thus work to make AI progress.

Security

- Security: the safety and security of AI should be tested and assured across the entire life cycle within an explicit and well-defined domain of use. In addition, AI should safeguard the people who are impacted.

- Beyond the definition:

  - This notion is connected to robustness and privacy. Securing AI is key for people to trust it.

  - The idea is not only to secure the systems, but also to secure the people, their data and the way the AI interacts with them, especially when dealing with high-risk AI domains.

Next steps ? Suggestions

❑Short-Term

- Liaise with the tools group to review the Principles
- Present the work to Trusted AI Committee 2020-11-19
- Present the work to the Board 2020-12-01
- Principles Working Group will publish a blog announcing the Principles

❑Mid-Term

- Call for Volunteer LF-AI Projects: examine and adopt the Principles – at various stages in the life-cycle
- Take one specific LF-AI project or LF project and test directly the implementation of these principles and guidelines all along the lifecycle
- Share the results within the wider community of LF and LF-AI and Data
- Communication : Blogs, Webinars, Conference submissions
- Assess the relationship of the Principles with existing and emerging trusted AI toolkits and software

❑Long-Term

- Training & Communication Integration :
  - Include Principles in future LFAI communication
  - Include the Principles in LF-AI Ethics course
- Explore:
  - Coaching Methods based our guidelines
  - Methods of audits
  - Badging or Certificates

›

References and Resources

- LF-AI https://lfai.foundation/
- LF-AI Committee link https://wiki.lfai.foundation/
- LF-AI Trusted AI Committee https://wiki.lfai.foundation/display/DL/Trusted+AI+Committee

References and Resources

- [ACM] ACM Principles for Algorithmic Transparency and Accountability https://www.acm.org/binaries/content/assets/publicpolicy/2017_usacm_statement_algorithms.pdf
- [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
- [EUFeb2020] 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
- [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
- [OECD] Organisation for Economic Co-operation and Development https://www.oecd.org/going-digital/ai/principles/
- [SoA] State of the Art: Reproducibility in Artificial Intelligence Odd Erik Gundersen, Sigbjørn Kjensmo, Department of Computer Science Norwegian University of Science and Technology https://www.researchgate.net/publication/326450530_State_of_the_Art_Reproducibility_in_Artificial_Intelligence

- Appendix
- More on Next Steps

Consider for next steps ?
A few proposals to be tested from an operational perspective

- Take one specific LFAI project or LF project and test directly the implementation of these principles and guidelines

- Include the definitions in the machine, then second define process for the AI to always follow the principles and raise hands when there is contradiction, issues with these principles. It mean to define processes

- Begin with the data used, how they were collected, how they will be used, the objectives and how it can improve people's life opposed to how it may harm people. Check hardware, software, implement principles. How? Begin with the objectives of the AI (do good, no harm) check the data used, how they are collected, why, for how long and how it is secured, check biases in algorithms, are they needed, why, should we correct them ?

- Identify/assess potential risks. Identify how to handle them

- Auditing the machine is an additional way to proceed and check how AI is working, correct biases when relevant, or add biases when relevant

Consider for Next steps ?
A few proposals to be tested from an operational perspective

- Appoint a responsible for each step and for the global process of learning: who is responsible for the way AI is working ? How to address accountability for successive AI during their life-cycle ?

- Use the different steps of AI life cycle

- Audit at all stages. At the stage of the equipment and software, of implementation, check how the data is handled by the machine, see if new biases appears, if they are needed, how to correct them

- Again at the stage of the run, check data, interaction between data, new data, biases, correction needed

- At the stage of the run and maintenance, check, repair

- At all stage, audit, checking with process that the different principles are applied. Apply corrections when needed. Certify

- Last but not least, in terms of ethics, shouldn't make mandatory the application of the common principles shared or at least make it a standard for AI to be acknowledged and trusted