LFAI & Data

LF AI & Data Day EU
Introducing RREPEATS: The Trusted AI Principles

Trusted AI Committee - Principles working group

June 10, 2021

LF AI & DATA

Introduction: agenda & first interactions

Agenda

20 minutes presentation
5 minutes Q/A

First interactions

Trusted AI ?

- What are the first words that come to your mind when talking about trusted AI ? Why ?

Trusted AI: Motivations & Overview

- What are the first words that come to your mind when talking about trusted AI ?
- Why ?
- Having common principles is not so easy
- Depending where you live, even if the principles are shared, the concepts are not always the same and the way they are applied or prioritized may be different
- To encourage a common approach across the globe, despite regional differences, we emphasize:
  - No principle is of higher priority than another

The 8 LFAI Principles for Trusted AI – (R)REPEATS

Reproducibility	Robustness	Equitability	Privacy
Explainability	Accountability	Transparency	Security

The principles are of equal importance and value.

No principle is of higher priority than another.

The principles are related to each other.

LF AI & DATA

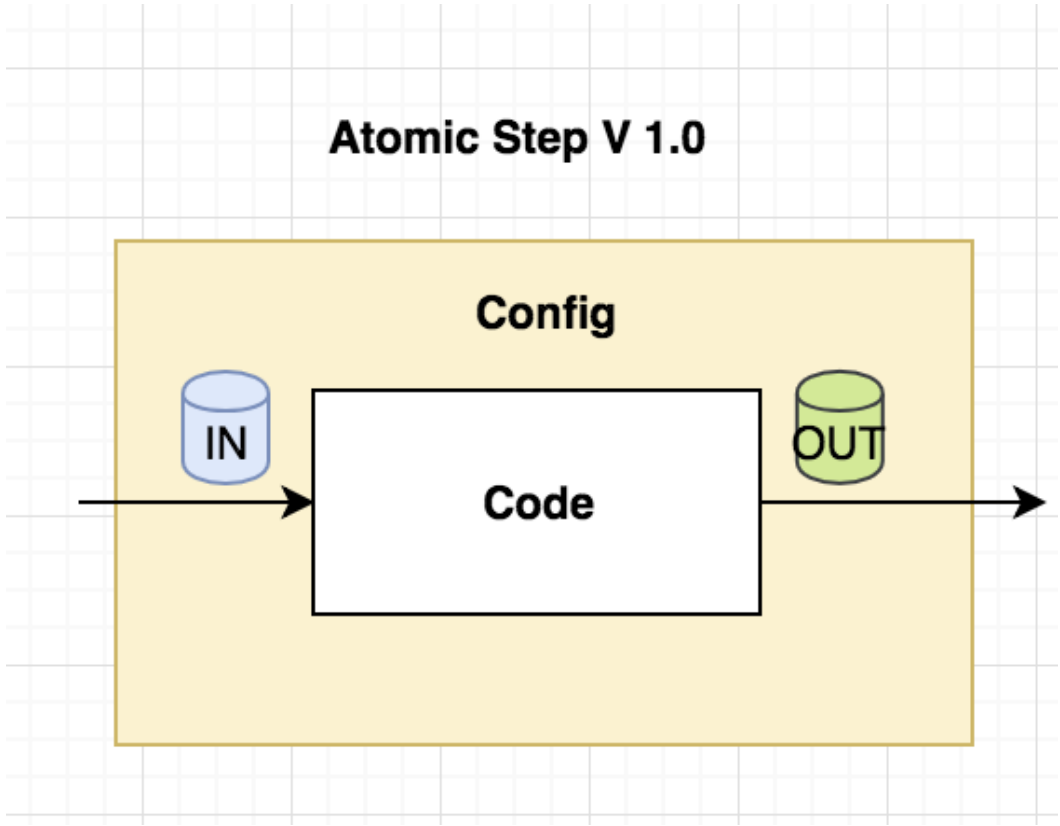
Why are the Principles important

- 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

Illustration through a few principles

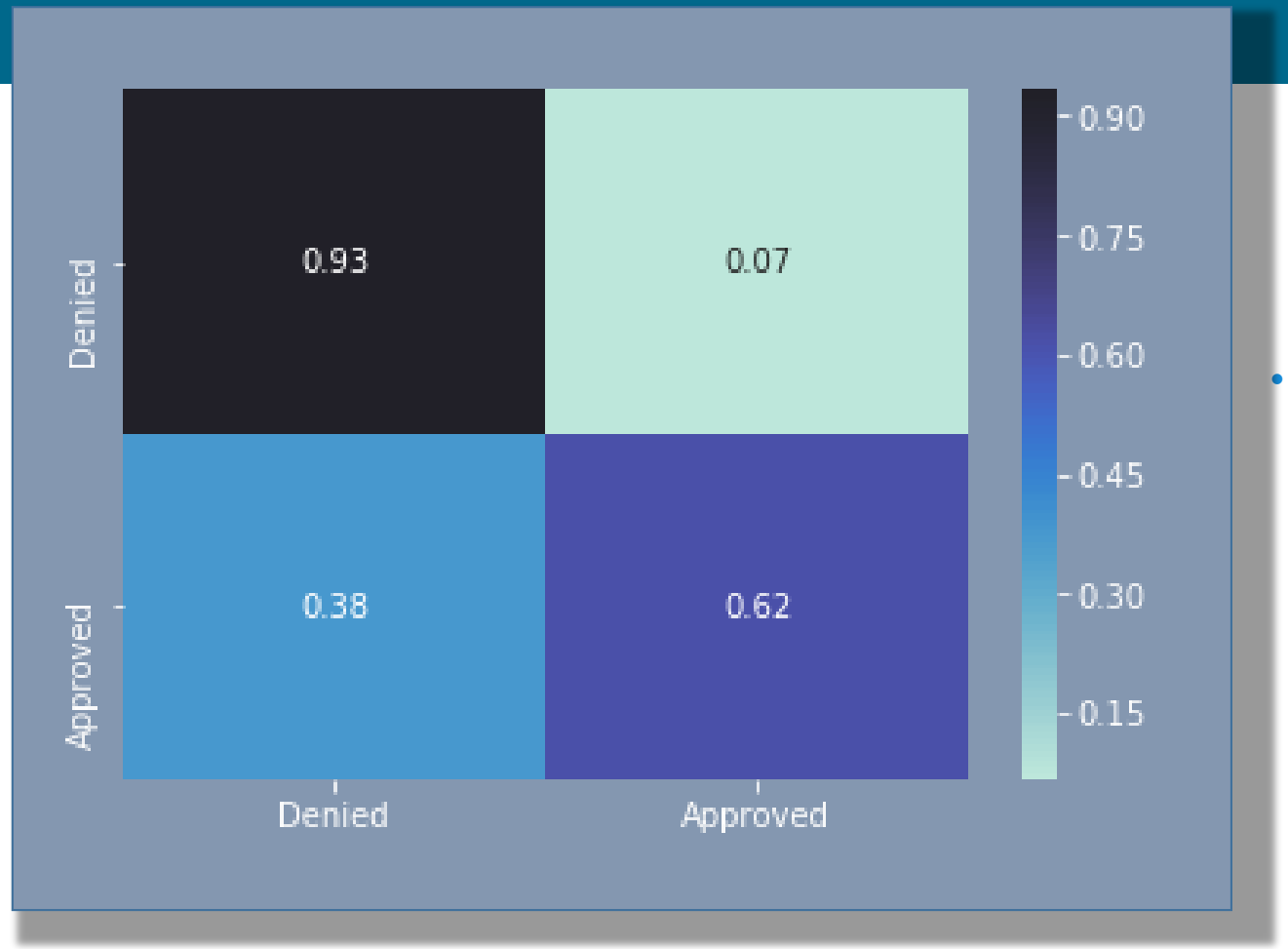
- There is a complete document describing the principles, their definitions and the next steps. We encourage you to read and use them
- In this presentation we will examine the principles to illustrate the work that has been completed and provide background for the definitions.

Reproducibility



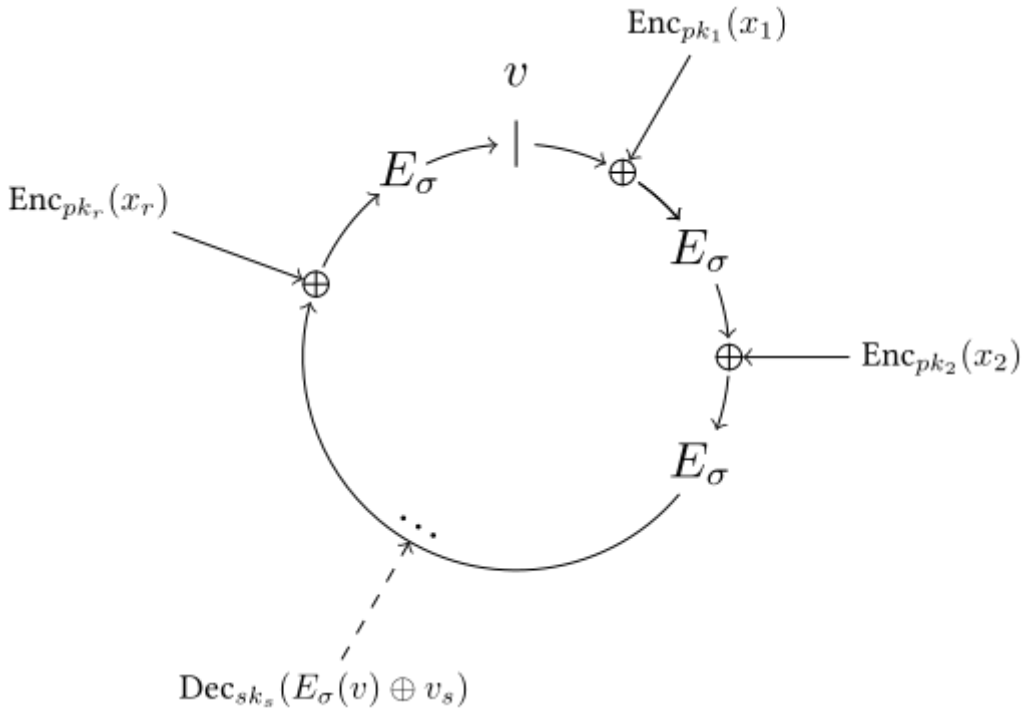
- **Reproducibility** is the ability of an independent team to replicate in an equivalent AI environment, domain or area, the same experiences or results using the same AI methods, data, software, codes, algorithms, models, and documentation, to reach the same conclusions as the original research or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

Equitability



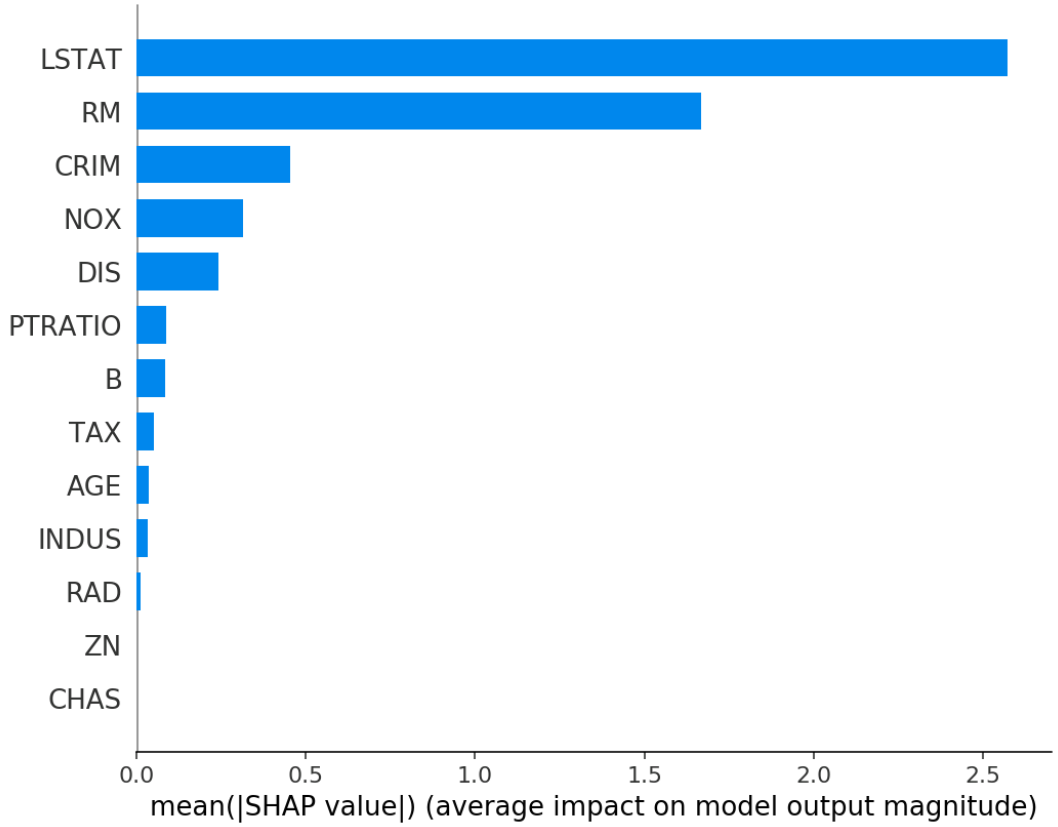
- **Equitability** for AI and the people behind AI should take deliberate steps - in the AI life-cycle - to avoid intended or unintended bias and unfairness that would inadvertently cause harm.

Privacy



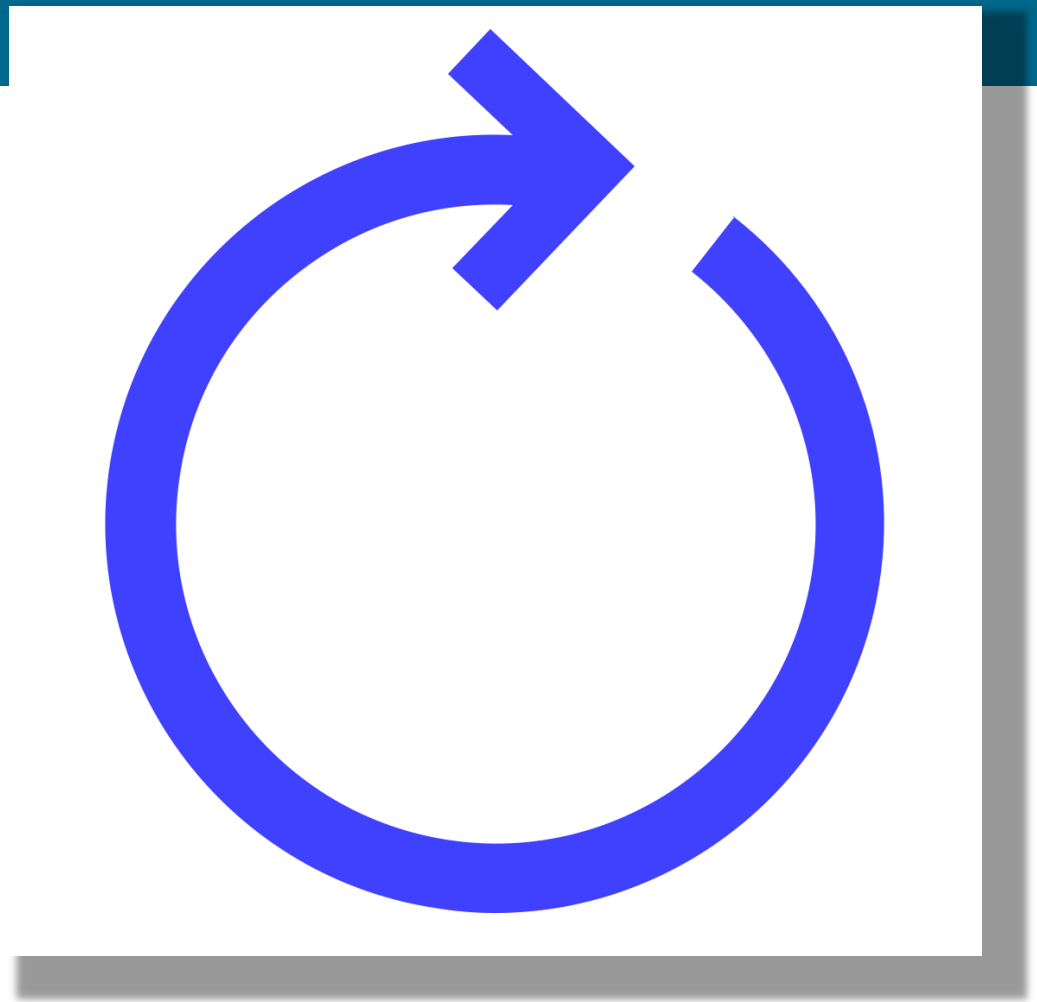
- **Privacy** requires AI systems to guarantee privacy and data protection throughout a system's entire lifecycle. The lifecycle activities include the information initially collected from users, as well as information generated about users throughout their interaction with the system e.g., outputs that are AI-generated for specific users or how users responded to recommendations.

Explainability



- **Explainability** is the ability to describe how AI works, i.e., makes decisions. 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.

Robustness

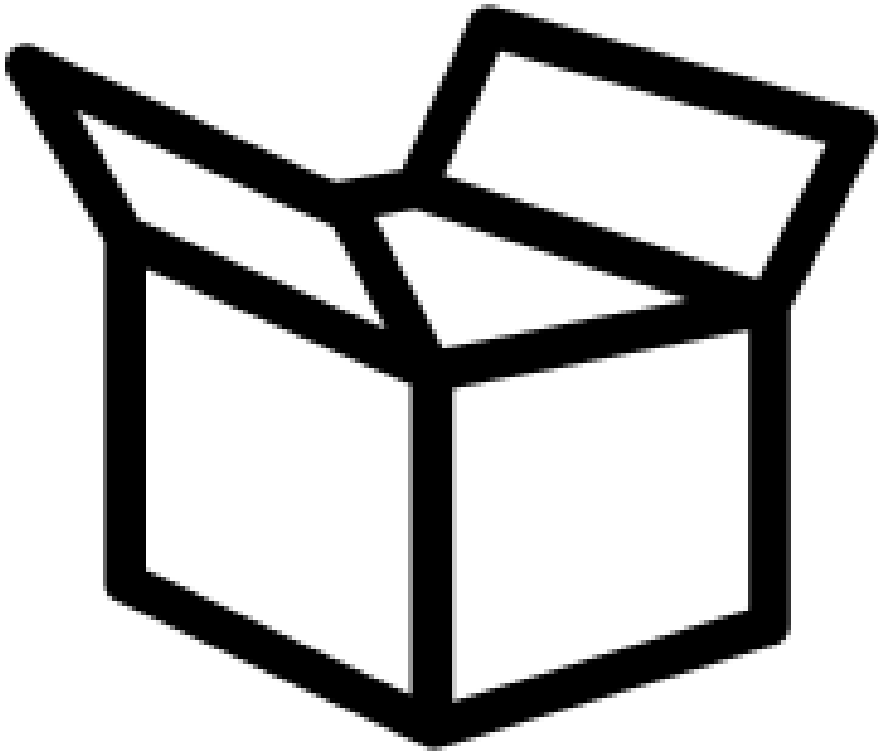


- **Robustness** refers to the stability, resilience, and performance of the systems and machines dealing with changing ecosystems. AI must function robustly throughout its life cycle and potential risks should be continually assessed and managed.

Accountability

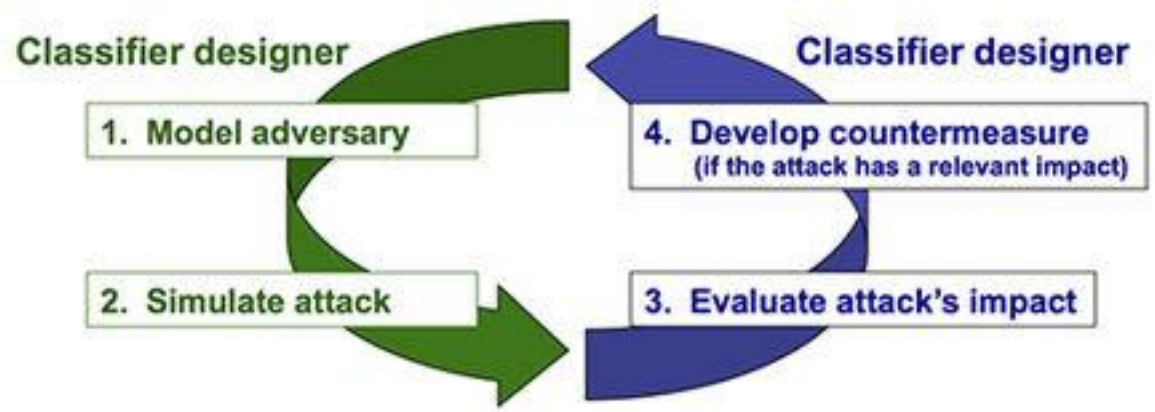
- **Accountability** requires AI and people behind the AI to explain, justify, and take responsibility for any decision and action made by the AI. Mechanisms, such as governance and tools, are necessary to achieve accountability.

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.

Security



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

Conclusion

- Trust is the basis of any human activity,
- Trust is one of the foundations for healthy human relationships
- Without trust, little can be accomplished. Achieving shared goals is more difficult
  - To provide trust in AI: the challenge is to include ways of analyzing what is good and safe, and what is evil and unsafe
- AI is just another tool that humans shape to suit their needs in compliance with their world.
  - Think about the revolution when the first knife was made.
    - A knife can kill, it can also cut vegetables and be used to build.
    - The evil is always behind the one using the tool.
    - The idea is how to master/manage this tool.
- Call for volunteers to test the Trusted-AI Principles

References and Resources

- [LF-AI] The Trusted-AI Principles document bit.ly/lfai-trustedai-principles
- [LF-AI Blog] LF AI & Data Announces Principles for Trusted AI
- [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

Contributions

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 / Seldon)
- Marta Ziosi (AI for People)
- Haluk Demirkan (Milgard School of Business – University of Washington)

Thank you

› ANNEX

Actions realized in 2020/2021

- Principles document drafted and shared
  - Liaise with the tools group to review the Principles
  - Present the work to Trusted AI Committee
  - Present the work to the Board 12/03/2020
- Principles Group published a Blog Announcing the Principles
- Communication : Blogs, Webinars, Conference submissions
  - ✓ Blog drafted and available on the LFAI website https://lfaidata.foundation/blog/2021/03/09/rrepeats-an-introduction-to-the-principles-for-trusted-ai-on-10-february-2021
  - ✓ First webinar: Introducing RREPEATS realized the 10th of February 2021 - 18 people attended, session recorded
  - ✓ Second webinar: The Trusted AI Principles - Practical Examples 28th of April 2021 – session recorder and available on YouTube The Trusted AI Principles - Practical Examples – YouTube & Blog https://lfaidata.foundation/blog/2021/05/13/trusted-ai-principles-rrepeats-practical-examples-review/ - Session recording / Session slides
  - ✓ Mapping of principles tools first draft
  - ✓ The main project page is here https://wiki.lfaidata.foundation/display/DL/Principles+Working+Group

Next steps and actions for 2021

- 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

  A new webinar is planned in April 2021, more focused on showing use cases, other webinars to present RREPEATS to be planned
- Assess the relationship of the Principles with existing and emerging trusted AI toolkits and software
- 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