Funded PhD Candidate position 2024-2027

- **Title** : Supporting Trustworthy AI with Enterprise Blockchain
- **Thesis Director**: Pr. Camille Salinesi
- **Thesis co-Director**: Dr. Nicolas HERBAUT
- **Keywords**: blockchain, AI, trust

1 Context

1.1 Research

The following thesis proposal is a current research field in the Centre de Recherche en Informatique de l'université de Paris 1 Panthéon-Sorbonne, and builds around past and present research effort in Entreprise Blockchain and Artificial Intelligence

1.2 Motivation

Trustworthy artificial intelligence (AI) is a multifaceted concept underpinned by three critical pillars: Trustworthy Machine Learning, Trustworthy Inference (often associated with Explainable AI, or XAI), and Trustworthy Data Management [1]. Together, these components form the backbone of ethical and reliable AI systems. Trustworthy Machine Learning ensures that AI models are developed with integrity, fairness, and accountability. Trustworthy Inference focuses on the ability of AI systems to provide transparent and understandable explanations for their decisions, thereby enhancing user trust. Lastly, Trustworthy Data emphasizes the importance of high-quality, secure, and privacy-respecting data as the foundation upon which AI models are trained. The interconnection of these concepts is essential for the realization of AI systems that are not only effective but also align with ethical standards and societal values.

We will see throught this section that blockchain may be an interesting solution to support trustworthy AI.

1.2.1 Trustworthy Machine Learning

The primary obstacle to achieving trustworthy learning is the potential compromise of privacy during the model's operation. The Introduced by Google in 2016, Federated Learning (FL) proposes a collaborative model training approach across multiple devices without necessitating the sharing of private data, under the coordination of a centralized server [2]. This methodology holds significant promise for sensitive sectors such as healthcare and finance, where the sharing of private information poses substantial risks. Despite FL's potential for preserving data privacy, it remains susceptible to similar vulnerabilities as conventional machine learning models, including targeted data poisoning attacks by malicious actors aiming to compromise the global model with erroneous updates [3].

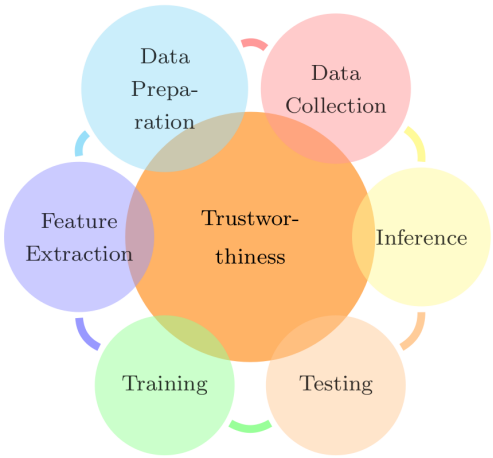

Figure 1: chain of trust [1]

The integration of blockchain technology with FL has emerged as a compelling advancement [2], drawing notable interest for its potential to enhance FL's efficacy through innovative solutions. This synergy aims to address the unique requirements of various applications, presenting theoretical and practical improvements for FL's performance.

1.2.2 Trustworthy AI Data

The process of preparing data for AI models, encompassing data design, sculpting (including cleaning, valuation, and annotation), and evaluation, is pivotal in shaping the reliability and trustworthiness of AI systems. These stages involve meticulous manual efforts, highlighting the importance of careful data management to foster dependable AI solutions [4].

Blockchain technology, with its immutable ledger capabilities, offers a method for documenting, executing, and auditing data preparation processes, ensuring that only trustworthy operations are carried out.

1.2.3 Trustworthy AI Inteference

Furthermore, the role of explainable AI (XAI) cannot be overstated in fostering the acceptance and credibility of AI outcomes. The integration of blockchain technology, leveraging mechanisms such as smart contracts, trusted oracles, and decentralized storage, presents innovative pathways to achieve explainability and trust in AI systems [5].

In summary, the interplay between data integrity, privacy, federated learning, blockchain technology, and explainable AI forms a complex ecosystem that underscores the importance of each element in the development of trustworthy and efficient AI systems.

1.2.4 Trustworkthy AI Ecosystem

Emphasizing the significance of securing both data and algorithms comprehensively through a systemic approach is crucial. Trust in individual components alone is

insufficient if the overall system remains vulnerable.

From a socio-economic viewpoint, distributing the benefits of AI across various stakeholders is essential for fostering an environment conducive to mutual collaboration.

In our previous publication, "SAIaaS: A Blockchain-based Solution for Secure Artificial Intelligence as-a-Service" [6], we explored this concept further and extended our analysis of trustworthy digital ecosystems in our article in press [7].

2 Research Problems

The pursuit of Trustworthy AI encompasses the intricate challenge of ensuring that artificial intelligence systems are developed and operate within ethical guidelines, are transparent in their decision-making processes, and are built upon secure and privacy-respecting foundations. This endeavor presents several research problems that are critical to address:

- How to use blockchain to build a trustworthy data preparation pipeline for AI?
- How to use blockchain to support trustworthy machine learning?
- How to use blockchain to assure the exaplainability of AI inference?
- How to globally assemble trustworthy AI components to build a full trustworthy end-to-end AI solution?

3 State of the art for blockchain and trustworthy AI

In today's world, where technology lets us collect and use data from many places, it's really important to make sure this data is accurate and can be trusted. This helps people and organizations make smart decisions. One key way to do this is by checking where the data comes from and if it's reliable, as mentioned in [8]. Blockchain technology has been useful in several areas like vehicular networks [9], Big Data [10], and IOT [11] because it helps manage data from different places with varying levels of trust.

When it comes to using data for machine learning (ML), preparing the data properly is important to make sure it's used right. But, these steps can sometimes mess with the quality and trustworthiness of the data, especially when choosing what data to train and test the algorithms on. Problems can arise if these steps mix data from different provenance in ways that could break privacy rules or introduce bias, like when removing outlier.

> Our goal is to keep data trustworthy by (1) making sure we can always trace back to where our data came from and (2) ensuring that our methods of cleaning and organizing the data do not make it less reliable. Doing both of these things at the same time isn't very common, so we're focusing on gathering data carefully and categorizing how we process it. This way, we can make sure our data remains trustworthy.

The use of blockchain to enhance trust and privacy in machine learning and inference processes has been extensively explored, showing significant benefits in areas like privacy enhancement [12], fraud detection [13], and audit processes [14]. Among privacy-preserving techniques in artificial intelligence, Federated Learning and Hybrid Approaches [15] stand out. However, these methods are vulnerable to a variety of

attacks, and effective countermeasures are not always readily available, indicating the need for defensively designed blockchain solutions.

> Our objective is to investigate various countermeasures and blockchain designs that can support the implementation of privacy-preserving machine learning frameworks in a verifiably trustworthy manner. To this aim, Blockchain software patterns offers resonable solutions to concrete problems and should be put to work to solve the specific problems faced by such developments.

Explainable AI, utilizing blockchain, has recently been introduced in various fields, such as agriculture [16], healthcare [17], and cybersecurity [18]. However, this area faces several significant challenges, including security vulnerabilities, a lack of interoperability for explanations, and issues with fairness [19]. Concurrently, advances in blockchain semantic interoperability, as seen in Horizon Europe projects like OntoChain [20], enable the use of blockchain as a comprehensive, trusted knowledge base and reference for AI explainability.

> Building on these developments, our goal is to make several contributions that broadly enhance trustworthiness attributes (such as those mentioned above) through the unique capabilities of blockchain.

4 Thesis Organization

The work to be done will be as follows:

1. Write a survey paper on trustworthy AI supported by blockchain, ensuring the state of the art remains updated throughout the PhD.
2. Develop models to bolster trustworthy AI at each stage of the trust chain. Each model should clearly outline the requirements for such systems, drawing on the survey to address outstanding issues in the current state of the art and provide strong evidence that trustworthiness requirements are met.
    1. Data: A model to support auditable data provenance, fusion, and processing.
    2. Learning: A model to establish best practices in Blockchain-aided privacy-preserving ML solutions, potentially enhanced by ML techniques.
    3. Inference: An end-to-end model to ensure trustworthy and actionable AI explainability.
    4. Ecosystem: model and propose end-to-end solution for a sustainable trustworthy BC-aided AI ecosystem
3. Implement and demonstrate the viability of these models through experiments, ideally based on industrial cases.
4. Propose a comprehensive end-to-end approach for trustworthy AI.
5. Disseminate research findings through scientific articles, patents, publicly accessible open-source tools, and technology transfer.

5 Candidate

A good candidate should have a master or engineering degree in computer Science or Information System. She/He should have good programming skills and be knowledgeable in the blockchain ecosystem, AI, software engineering and be fluent in English. We do value candidates with a result-oriented mindset, with previous industrial or research experience.

References

[1] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. Van Moorsel, "The relationship between trust in AI and trustworthy machine learning technologies," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 272–283.

[2] P. M. Mammen, "Federated learning: Opportunities and challenges," *arXiv preprint arXiv:2101.05428*, 2021.

[3] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in *Computer security–ESORICS 2020: 25th european symposium on research in computer security, ESORICS 2020, guildford, UK, september 14–18, 2020, proceedings, part i 25*, 2020, pp. 480–501.

[4] W. Liang, G. A. Tadesse, D. Ho, L. Fei-Fei, M. Zaharia, C. Zhang, and J. Zou, "Advances, challenges and opportunities in creating data for trustworthy AI," *Nature Machine Intelligence*, vol. 4, no. 8, pp. 669–677, 2022.

[5] M. Nassar, K. Salah, M. H. ur Rehman, and D. Svetinovic, "Blockchain for explainable and trustworthy artificial intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 1, p. e1340, 2020.

[6] N. Six, A. Perrichon-Chrétien, and N. Herbaut, "Saiaas: A blockchain-based solution for secure artificial intelligence as-a-service," in *The international conference on deep learning, big data and blockchain (deep-BDB 2021)*, 2022, pp. 67–74.

[7] Y. Ding and N. Herbaut, "A conceptual model for blockchain-based trust in digital ecosystems," *Acta Cybernetica*, Jun. 2024.

[8] C. Dai, D. Lin, E. Bertino, and M. Kantarcioglu, "An approach to evaluate data trustworthiness based on data provenance," in *Secure data management: 5th VLDB workshop, SDM 2008, auckland, new zealand, august 24, 2008. Proceedings 5*, 2008, pp. 82–98.

[9] H. Xu, S. Qi, Y. Qi, W. Wei, and N. Xiong, "Secure and lightweight blockchain-based truthful data trading for real-time vehicular crowdsensing," *ACM Transactions on Embedded Computing Systems*, vol. 23, no. 1, pp. 1–31, Jan. 2024.

[10] K. Mehboob Khan, W. Haider, N. Ahmed Khan, and D. Saleem, "Big data provenance using blockchain for qualitative analytics via machine learning " *JUCS - Journal of Universal Computer Science*, vol. 29, no. 5, pp. 446–469, May 2023.

[11] C. A. Ardagna, R. Asal, E. Damiani, N. E. Ioini, M. Elahi, and C. Pahl, "From trustworthy data to trustworthy IoT: A data collection methodology based on blockchain," *ACM Transactions on Cyber-Physical Systems*, vol. 5, no. 1, pp. 1–26, Dec. 2020.

[12] P. Kumar, R. Kumar, M. Aloqaily, and A. K. M. N. Islam, "Explainable AI and blockchain for metaverse: A security and privacy perspective," *IEEE Consumer Electronics Magazine*, vol. 13, no. 3, pp. 90–97, May 2024.

[13] Z. Jovanovic, Z. Hou, K. Biswas, and V. Muthukkumarasamy, "Robust integration of blockchain and explainable federated learning for automated credit scoring," *Computer Networks*, vol. 243, p. 110303, Apr. 2024.

[14] S. Sachan and X. Liu (Lisa), "Blockchain-based auditing of legal decisions supported by explainable AI and generative AI tools," *Engineering Applications of Artificial Intelligence*, vol. 129, p. 107666, Mar. 2024.

[15] N. Khalid, A. Qayyum, M. Bilal, A. Al-Fuqaha, and J. Qadir, "Privacy-preserving artificial intelligence in healthcare: Techniques and applications," *Computers in Biology and Medicine*, p. 106848, 2023.

[16] H.-Y. Chen, K. Sharma, C. Sharma, and S. Sharma, "Integrating explainable artificial intelligence and blockchain to smart agriculture: Research prospects for decision making and improved security," *Smart Agricultural Technology*, vol. 6, p. 100350, Dec. 2023.

[17] A. S. Albahri, A. M. Duhaim, M. A. Fadhel, A. Alnoor, N. S. Baqer, L. Alzubaidi, O. S. Albahri, A. H. Alamoodi, J. Bai, A. Salhi, J. Santamaría, C. Ouyang, A. Gupta, Y. Gu, and M. Deveci, "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," *Information Fusion*, vol. 96, pp. 156–191, Aug. 2023.

[18] S. Benedict, "EA-POT: An explainable AI assisted blockchain framework for HoneyPot IP predictions," *Acta Cybernetica*, vol. 26, no. 2, pp. 149–173, Nov. 2022.

[19] A. Rawal, J. McCoy, D. B. Rawat, B. M. Sadler, and R. S. Amant, "Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 6, pp. 852–866, 2021.

[20] T. G. Papaioannou, V. Stankovski, P. Kochovski, A. Simonet-Boulogne, C. Barelle, A. Ciaramella, M. Ciaramella, and G. D. Stamoulis, "A new blockchain ecosystem for trusted, traceable and transparent ontological knowledge management: Position paper," in *Economics of grids, clouds, systems, and services: 18th international conference, GECON 2021, virtual event, september 21–23, 2021, proceedings 18*, 2021, pp. 93–105.