Assessing the Sustainability and Trustworthiness of Federated Learning Models

Chao Feng*[a], Alberto Huertas Celdrán[a,b], Pedro Miguel Sánchez Sánchez[b], Lynn Zumtaugwald[a], Gérôme Bovet[c] and Burkhard Stiller[a]

[a] Communication Systems Group CSG, Department of Informatics IfI, University of Zurich UZH, 8050 Zürich, Switzerland
[b] Department of Information and Communications Engineering, University of Murcia, 30100 Murcia, Spain
[c] Cyber-Defence Campus within Armasuisse Science & Technology, 3602 Thun, Switzerland

ARTICLE INFO

Keywords:
Sustainable AI
Carbon Footprint
Federated Learning
Distributed Machine Learning

Abstract

Artificial intelligence is widely used in various sectors and significantly impacts decision-making processes. Novel AI paradigms, such as Federated Learning (FL), focus on training AI models collaboratively while preserving data privacy. In such a context, the European Commission's AI-HLEG group has highlighted the importance of sustainable AI for trustworthy AI. While existing literature offers several solutions for assessing the trustworthiness of FL models, a significant gap exists in considering sustainability associated with FL. Thus, this work introduces the sustainability pillar to the trustworthy FL taxonomy, making this work the first to address all AI-HLEG requirements. The sustainability pillar assesses the FL system's environmental impact, incorporating notions and metrics for hardware efficiency, federation complexity, and energy grid carbon intensity. An algorithm is developed to evaluate the trustworthiness of FL models, incorporating sustainability considerations. Extensive evaluations with the FederatedScope framework and various scenarios demonstrate the effectiveness of the proposed solution.

1. Introduction

Over the past few decades, Artificial Intelligence (AI) has undergone pervasive integration into various facets of society, encompassing applications such as recreational gaming, disease diagnosis, text and art generation, or autonomous driving [6]. The relevance obtained by AI has amplified the necessity of sustainability, which traverses environmental, social, economic, and ethical dimensions. Delving into the specifics, utilizing Deep Learning (DL) models, predominantly characterized by resource-intensive computational demands during training and evaluation, leads to a significant carbon footprint. Simultaneously, DL systems heavily rely on massive data, and unsustainable data management methodologies incur superfluous energy consumption. Furthermore, ethical considerations assume paramount significance in sustainable AI, aiming to preclude negative repercussions, including bias, discrimination, and privacy infringements.

In conjunction with robustness, transparency, fairness, and accountability, sustainable AI assumes a central role in nurturing long-term societal acceptance and establishing trust in AI systems. As the adoption of AI technologies continues to proliferate across industries and impact many societal aspects, ensuring the trustworthiness of AI becomes paramount. In this direction, governing bodies and regulatory authorities worldwide recognize the necessity of addressing trustworthy AI [9]. For instance, the High-level Expert Group on Artificial Intelligence (AI-HLEG [8]) in Europe has played a pivotal role in formulating legal frameworks and guidelines designed to shape and oversee the development of trustworthy AI [23]. In a more granular context, the AI-HLEG defines seven prerequisites for trustworthy AI, which are: 1) human agency and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) fairness, 6) environmental well-being, and 7) accountability.

As highlighted by the AI-HLEG, data privacy is a challenging and active research topic within trustworthy AI. In 2016, Google introduced Federated Learning (FL) [15], an innovative paradigm that enables multiple clients to collaboratively train models without necessitating the exchange of private data. Nowadays, FL confronts multifaceted challenges, spanning scalability, single point of failure, architectural design, or privacy and security concerns, among others [2]. However, while FL inherently incorporates privacy-preserving features, trustworthy AI remains a pivotal dimension within FL systems.

In this context, prior works [3, 25] defined a baseline by formulating taxonomies for trustworthy ML, DL, and FL. Other works, such as [20], implemented algorithms and frameworks for assessing the trustworthiness of FL systems. However, environmental well-being is completely missing in those works. More in detail, Carbon dioxide equivalent (CO_2eq), a unit based on the global warming potential (GWP) of different greenhouse gases, has not been considered while assessing the FL trustworthiness, as articulated by AI-HLEG. In this sense, hardware efficiency, federation complexity, or energy grid carbon intensity should be considered and studied while assessing the trustworthiness of FL to raise awareness and design optimum federation configurations.

To improve the previous challenges, the main contributions of this work are:

- The review of the state of the art regarding sustainable and trustworthy AI. As a result, it has been designed a

*Corresponding author. Email address: cfeng@ifi.uzh.ch (C. Feng)
ORCID(s): 0000-0002-0672-1090 (C. Feng*); 0000-0001-7125-1710 (A. Huertas Celdrán); 0000-0002-6444-2102 (P.M. Sánchez Sánchez); 0000-0002-4534-3483 (G. Bovet); 0000-0002-7461-7463 (B. Stiller)

novel Trustworthy FL taxonomy composed of seven pillars (privacy, robustness, fairness, accountability, federation, explainability, and sustainability). The sustainability pillar is novel, and it is composed of three notions (carbon intensity, hardware efficiency, and federation complexity) and ten metrics.

- The design and implementation of an algorithm to evaluate the sustainability and trustworthiness of FL models (source code available in [28]). The proposed algorithm improves related work in implementing metrics assessing the sustainability of FL models. In particular, three notions and ten metrics have been proposed for FL sustainability computation, considering the CO_2eq impact of heterogeneous FL models. The algorithm combines the ten sustainability metrics with 41 already proposed in the literature for the remaining six trustworthy FL pillars to give an overall score of trustworthy AI.
- The deployment of the algorithm in a real FL framework, called FederatedScope [26], and the evaluation of its performance in different scenarios with several configurations in terms of hardware efficiency, federation complexity, and carbon-intensity of energy grids. The obtained results demonstrated the suitability of the framework while considering sustainability as another factor to measure the trustworthiness of FL.

The remainder of this paper is structured as follows. Section 2 contains findings from the literature review on trustworthiness and sustainability in FL. Section 3 presents a detailed analysis of the sustainability pillar and its metrics. Section 4 presents the design and implementation details of the proposed algorithm. Section 5 validates the algorithm in a use case and presents the results of the performed experiments. Section 6 discusses the current limitations in sustainability computation for FL. Finally, Section 7 provides conclusions and future work.

2. Related Work

This section reviews recent and relevant work done in the literature regarding trustworthy FL evaluation and carbon emission estimation for AI/FL-based computing.

2.1. Trustworthy FL

Table 1 summarizes the existing trustworthy FL taxonomies and their coverage of trustworthy FL pillars defined by the AI-HLEG. The taxonomy from Shi et al. [22] reviewed the issue of fairness in FL and its evaluation mechanisms. This study only covers the pillar of fairness and partially the federation one since it discusses fair client selection. Liu et al. [11] provided a taxonomy covering the pillar of privacy, robustness, and partially the pillar federation. Tariq et al. [25] proposed an architecture for FL trustworthiness. Its taxonomy covers privacy, fairness, explainability, and robustness pillars and includes requirements two, three, and five defined by the AI-HLEG. Zhang et al. [27] also surveyed trustworthy FL, but focusing on the legal aspects of security, privacy, and robustness pillars. The taxonomy that covers the most pillars and requirements defined by the AI-HLEG is the trustworthy FL taxonomy from Sánchez et al. [20]. The taxonomy contains the pillars i) privacy, ii) robustness, iii) fairness, iv) explainability, v) accountability, and vi) federation. For each pillar, notions and metrics are defined. In total, 36 metrics are defined that can be used to evaluate the trustworthiness score of a given FL system.

After reviewing the literature, the most important limitation becomes present when comparing the taxonomy to the requirements defined by the AI-HLEG and the existing taxonomies. The environmental impact of an FL system is not considered in the taxonomy, but environmental well-being has clearly been defined as one of the seven requirements for trustworthy AI by governing bodies [8]. Since [20] is the most advanced taxonomy that covers six of the seven requirements defined by the AI-HLEG, it is employed as the basis for extension considering the environmental impact of the system.

2.2. Sustainable AI/FL

Most works focus on estimating the carbon emissions of specific ML/DL models. Lucconi et al. [12] provided a survey on aspects that influence the CO_2eq of ML. Strubell et al. [24] estimated the financial and environmental costs of large natural language processing (NLP) models by analyzing the training and fine-tuning process. Luccioni et al. [13] estimated the carbon emissions of the large language model BLOOM having 176 billion parameters to be 50.5 tonnes of CO_2eq emission. Patterson et al. [17] estimated the energy consumption and computed the carbon emissions of the language models T5, Meena, GShard, Switch Transformer, and GPT-3 and highlighted opportunities to improve energy efficiency and CO_2eq emission such as sparsely activated DNNs and using energy grids with low carbon intensity. While the mentioned works focus mainly on energy consumption, George et al. [5] point out that water consumption to cool large data- and server centers also contributes heavily to the environmental impact of AI models and estimated the water consumption needed to run Chat-GPT.

In the field of FL, Qui et al. [19] provided a first look into the carbon footprint of FL models by incorporating parameters that are special to FL and comparing the emissions produced by FL models vs. emissions produced by centralized ML models. They concluded that FL models could emit up to two orders of magnitude of CO_2eq if the data is not identically distributed, which is often the case in FL. Similarly to estimating the carbon emissions of AI/FL models, tools to track carbon emissions and apply standardized measurements for better comparison of model emissions were developed in [13]. CodeCarbon [4] and the Experimental Emissions Tracker [7] can be used to track emissions during the training process, while the ML CO_2eq Calculator [10] can be used to calculate the emissions after training.

Despite the effort and work done in this research field, to the best of our knowledge, no work directly considered

Table 1
Existing Trustworthy FL Taxonomies and Their Coverage of Pillars and AI-HLEG Requirements

Authors (Year)	*Pillars/AI-HLEG Requirements*						
	Privacy	Fairness	Robustness	Accountability	Explainability	Federation	Sustainability
	3. Privacy and data governance	5. Diversity, non-discrimination, and fairness	2. Technical robustness and safety	7. Accountability and auditability / 1. Human agency and oversight	4. Transparency including explainability	2. Technical robustness and safety / 5. Diversity, non-discrimination and fairness	6. Environmental well-being
Shi et al. [22] (2021)	No	Yes	No	No	No	Partially	No
Liu et al. [11] (2022)	Yes	No	Yes	No	No	Partially	No
Tariq et al. [25] (2023)	Yes	Yes	Yes	No	Yes	No	No
Zhang et al. [27] (2023)	Yes	No	Yes	No	No	No	No
Sanchez et al. [20] (2023)	Yes	Yes	Yes	Yes	Yes	Yes	No
Qui et al. [19] (2023)	No	No	No	No	No	No	Partially
Carbon Code [4] (2023)	No	No	No	No	No	No	Partially
This work	Yes	Yes	Yes	Yes	Yes	Yes	Yes

the carbon emissions related to FL setups. In other words, the impact of the number of clients, aggregation functions, or data distributions in the carbon emissions have not been not considered by related work. Besides, no solution has incorporated the emissions produced by FL models into trustworthy FL despite environmental well-being clearly being defined as one of the seven key requirements for trustworthy AI/FL by the AI-HLEG [8].

3. The Sustainability Pillar of Trustworthy FL

This section describes the notions and metrics that make up the sustainability pillar of trustworthy FL. This pillar includes the carbon intensity of the energy grid, the efficiency of the underlying hardware, and the complexity of the federation. Besides, this section describes the complete taxonomy generated after adding the sustainability pillar to the most recent and complete existing trustworthy FL taxonomy.

3.1. Carbon Intensity

The carbon intensity of electricity varies in different parts of the world depending on the energy mix used to produce electricity. The United Nations Intergovernmental Panel on Climate Change (IPCC) [21] has provided a median value of grams of CO_2eq per kWh for different energy fuels. Wind and nuclear emit the least CO_2eq, with 12g and 11g of CO_2eq per kWh, and coal the most, with 820g of CO_2eq per kWh. Thus, an FL system that has used 500 kWh of energy to be trained would have emitted 5.5 kg of CO_2eq if it were trained on electricity produced by nuclear power and 410 kg of CO_2eq if it were trained on electricity produced by coal only. This showcases that the energy grid used to train FL systems plays a huge role in the carbon emissions produced. Similarly, the carbon intensity of the energy grid of countries varies by a remarkable factor. British Petroleum has published in their annual review of the world energy statistics [18] that the least carbon-intensive energy grid is used by the African country Lesotho with 20g of CO_2eq per kWh, and the most carbon-intensive energy grid is used by the South African country Botswana with 795 of CO_2eq per kWh in 2022.

Therefore, this notion seeks to measure the carbon impact of FL according to the following two metrics.

- **Client/Server Carbon Intensity**. These two metrics measure the carbon intensity of the energy grid utilized in the FL process from the perspectives of both the clients and the server. The value of these two metrics ranges from 20g of CO_2eq to 795 of CO_2eq by looking at the countries' energy grids, according to the IPCC report [21]. Theoretically, with the energy sources available today, the lowest possible energy grid would have 11g of CO_2eq per kWh only using wind energy and the highest possible 820g of CO_2eq only using coal energy. The energy grids used by clients can be determined by the location of the federation clients (retrieved from the IP address). The carbon intensity of the energy grid utilized by clients is determined by calculating the average of all the carbon intensities. For the carbon intensity of the energy grid used by the server, the energy grid of the country the server operates in is taken. Equation 1 illustrates the calculation process of this metric.

$$T_{Intensity} = S_{Intensity} + \frac{1}{n}\sum_{i=1}^{n} C_{nIntensity} \qquad (1)$$

Where $T_{Intensity}$ represents the total grid carbon energy intensity, $S_{Intensity}$ represents the server grid carbon intensity, and $C_{nIntensity}$ represents the grid carbon intensity of each client n.

3.2. Hardware Efficiency

The second notion that significantly impacts the energy consumption and, thus, the emissions of an FL system is the efficiency of the underlying hardware. Efficient hardware consumes less power to perform computational tasks. Lower power consumption translates to reduced energy requirements, leading to lower CO_2eq emissions. On the contrary, inefficient hardware generates more heat, necessitating additional cooling mechanisms, such as air conditioning or fans, that contribute to more CO_2eq emissions [10]. In FL systems, both the process of training local models and the aggregation of these models globally require heavy computational resources. Thus, the efficiency of the underlying hardware plays a significant role in the emissions produced by the FL system.

The performance of CPUs and GPUs can be described by different metrics, such as clock speed, Floating-Point Operations Per Second, or Instructions Per Second (IPS) [14]. It is important to note that none of these metrics provide a complete picture of the performance of the processing units, and different metrics are more relevant in certain use cases. Further, manufacturers of CPUs and GPUs often do not fully disclose the metrics of their products, which makes comparing them difficult. To solve this issue, lots of benchmarking software to evaluate the processor's performance across a range of tasks has been proposed. In terms of heat production of a processor, Thermal Design Power (TDP) is used as a specification in the industry [16]. It indicates the maximum amount of heat a computer component, such as a CPU or GPU, is expected to generate under normal operating conditions. TDP is typically expressed in watts and represents the maximum power consumption and heat dissipation expected under typical workloads. The smaller the number for TDP, the lower the power consumption of the processor. Therefore, the Hardware Efficiency notion proposes the following metrics.

- **Client/Server Hardware Efficiency**. To evaluate the efficiency of the underlying hardware in terms of computing power per unit of power consumed, it makes sense to divide the benchmark performance through the TDP, defining the power performance of the processor. A processor with a high power performance score is able to do a lot of computation with low energy consumption, and it is thus more efficient in terms of resource consumption [16]. It is measured in performance per Watt using Equation 2 and 3.

$$H_E = \frac{H_{BP}}{H_{TDP}} \qquad (2)$$

$$Total_E = S_E + \frac{1}{n}\sum_{i=1}^{n} C_{nE} \qquad (3)$$

Where H_E is the hardware efficiency score, H_{BP} is the hardware benchmark performance, H_{TDP} is the hardware TDP, S_E is the server hardware efficiency, and C_{nE} is the hardware efficiency of each client n.

3.3. Federation Complexity

The complexity and size of the federation impact the consumed energy and, thus, the emissions produced. Generally, the more complex the model, the higher the number of participants and the higher the energy consumption [19]. Therefore, the federation complexity notion considers the following metrics.

- **Number of Training Rounds**. This metric measures the number of federation training rounds. Each training round consumes energy for i) training the model on the client's side, ii) aggregating the model parameters on the server side, and iii) exchanging models between the client side and server side. Therefore, more training rounds emit more CO_2eq.

- **Dataset Size**. This metric measures the size of the dataset used by clients to train the FL models. Larger datasets need more computational resources regarding power, memory, and time to fit the model. Thus, larger datasets need more energy than smaller datasets and also produce more CO_2eq [10].

- **Model Size.** This metric measures the size of the model that is trained in the FL system. Large models typically require more computational resources and time to process each iteration, which results in higher energy consumption [10] at the client's side. Also, aggregating large models on the server side typically uses more energy than aggregating small models due to the number of weights. Furthermore, large models thus also introduce a communication overhead, again leading to more energy usage and CO_2eq emissions.

- **Number of Clients**. This metric measures the number of clients in the federation. The more clients participate in the federation, the more energy is used [19] for i) training, ii) aggregation, and iii) communication, and thus, the more CO_2eq are emitted.

- **Client Selection Rate.** This metric measures the client selection rate in the federation. Often, only a percentage of clients is selected per round [19]. The larger this percentage, the larger the communication overhead from the uplink communication, and the larger the CO_2eq emissions.

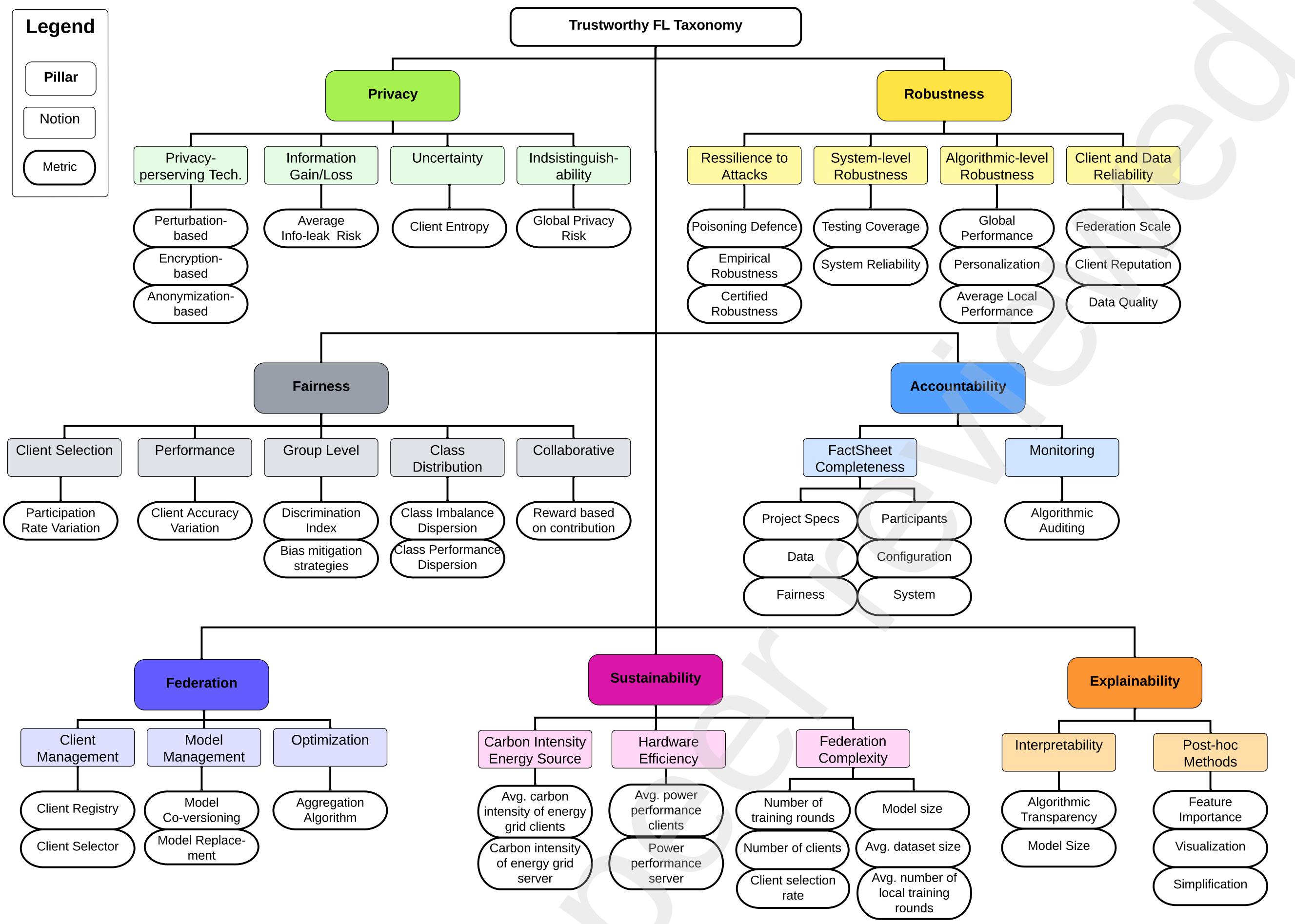

Figure 1: Trustworthy FL Taxonomy

- **Number of Local Training Rounds**. This metric measures the number of local training rounds within one federation training round. The higher the number of local training rounds, the higher the computational overhead on the client's side and the higher the energy consumption [10, 19].

3.4. Additional Pillars of Trustworthy FL

The six pillars defined by Sánchez et al. [20] together with the new one cover the seven requirements for trustworthy AI defined by the AI-HLEG [8] and constitute a comprehensive taxonomy. A visual representation of this taxonomy, including seven pillars, 23 notions, and 51 metrics, is presented in Figure 1. An overview of each pillar is given below. More detailed information can be found in [20].

3.4.1. Privacy

FL inherently provides a certain level of data privacy. However, it requires assumptions about the integrity of the various actors and entities within the federation. When participants are honest, but the aggregating server is 'honest-but-curious,' mechanisms to prevent information leakage are imperative. When all federation members exhibit 'honest-but-curious' behavior, the focus should shift to ensuring secure communication to prevent information leakage. Additionally, the potential for information leakage from external malicious attacks must be considered. To address these issues, this pillar considers four notions. The first emphasizes the adoption of privacy-preserving methods to enhance resilience against privacy attacks. The second notion involves metrics that quantify information gain or loss, considering the risk of information leakage inherent in the FL process. The final two notions relate to the probability of knowledge inference from client updates, necessitating a comprehensive and scientifically grounded approach to maintaining data privacy in FL models.

3.4.2. Robustness

Robustness in AI systems is imperative to safeguard against vulnerabilities to malicious applications and potential harm to humans. Within this context, existing literature delineates three distinct notions of robustness. The first notion, as highlighted in prior work, underscores the necessity for FL models to exhibit resilience against adversarial attacks, manifested through the introduction of perturbations or erroneous inputs. The second notion emphasizes the crucial need for robustness in both hardware and software utilized by participants in the training and deployment of FL models, a measure critical for thwarting cyberattacks. Finally, the third notion calls for reliability and robustness in the performance and customization of FL algorithms.

3.4.3. Fairness

Data-induced unfairness represents a significant challenge in AI, and this issue is particularly pronounced in FL due to the potential heterogeneity in the quantity and quality of data contributed by different clients. In this context, Client Selection Fairness emerges as the inaugural notion of this pillar, emphasizing the imperative for equitable participant inclusion. Beyond this, fairness in AI can be disaggregated into group-level and individual-level fairness. The former advocates for the absence of discrimination against any particular group, while the latter ensures equitable treatment of similar individuals, irrespective of their group affiliation. Transposing these fairness notions to FL, Group-level Fairness addresses disparities at the group level, whereas Performance Fairness and Class Distribution cater to individual-level fairness. Specifically, Performance Fairness ensures proportionality between a client's data contribution and their received rewards. Concurrently, Class Distribution scrutinizes label imbalances across the datasets of individual participants, ensuring a holistic approach to fairness in FL.

3.4.4. Explainability

AI guidelines stipulate the necessity for transparency across AI processes. Transparency within this context is frequently articulated as interpretability, a concept that is often erroneously equated with explainability. Interpretability is delineated as a model's inherent attribute that facilitates human understanding. Conversely, explainability pertains to the capacity to articulate the technical intricacies of AI systems. For models that are intrinsically interpretable, direct analysis can be enough for explanation. However, for models lacking this inherent interpretability, post-hoc methods, constituting the second notion of this pillar, become indispensable for enhancing their interpretability. In the realm of FL, where ML/DL models play a pivotal role in the training process, the imperative for explainability extends to the algorithmic model itself. Nonetheless, the imperative for data privacy in FL introduces complexities, as it restricts access to and analysis of raw data, necessitating innovative solutions to uphold explainability without compromising data privacy.

3.4.5. Accountability

Accountability stands as one of the seven imperative requirements for Trustworthy AI. The primary aspect of accountability is addressed through FactSheet Completeness [1]. IBM Research pioneered the concept of a FactSheet, a comprehensive document designed to meticulously record various facets of the entire ML/DL pipeline. Parallel to this, Monitoring emerges as another crucial notion of accountability. It underscores the responsibility of each participant to diligently verify that the FL models are constructed, developed, and deployed in strict alignment with the predetermined architectural and procedural guidelines. This ensures that despite the availability of comprehensive documentation, an active effort is made by all stakeholders to uphold the integrity and accountability of the FL models throughout their lifecycle.

3.4.6. Federation

The management of FL encompasses complex challenges pertaining to communication, efficiency, resource constraints, and security. Coordinating the learning processes across thousands of clients, while ensuring the integrity and security of the model, presents a formidable challenge. The convergence of global models may be impeded by data heterogeneity across clients, while inconsistencies in clients, networks, and limited resources may lead to client dropouts and training failures, adversely affecting the quality of the model. The critical notions within this pillar are identified as Client and Model Management, which delves into the administration of client and model information within the system, and Optimization Algorithm, which plays a pivotal role in influencing the model's performance and robustness.

4. Sustainable and Trustworthy FL Algorithm

This section provides the details of the algorithm in charge of assessing the sustainability and trustworthiness of FL models. The main contribution of this algorithm, compared to the literature, is the design and implementation of three notions and ten metrics dealing with the sustainability pillar and their integration with six other existing pillars (privacy, robustness, fairness, accountability, federation, and explainability). The following assumptions (A), functional requirement (FR), non-functional requirements (NF), and privacy constraint (PC) were considered during the algorithm design phase.

- A_1: The central server is honest. It is maintained by a trusted owner, and it does not interfere with the FL protocol maliciously.
- A_2: Clients of the federation are honest but curious. They trustfully report their metrics and statistics without maliciously interfering with the FL protocol.
- FR_1: The three notions and ten metrics of the Sustainability pillar must be represented in the algorithm. In addition, each of the remaining six trustworthy FL pillars must be considered, meaning that at least one metric from each pillar must be considered in the final score.
- FR_2: The final trustworthiness score must be a combination of the trustworthiness scores from all notions and pillars.
- NF_1: The algorithm should add minimal computation overhead and complexity to the server, participants, and FL model.
- NF_2: The algorithm should be modular and configurable.

Table 2
Metrics for Sustainability Pillar

Metric	*Description*	*Input*	*Output*	*Normalized Output*
		Notion: Carbon Intensity of Energy Source		
Avg, carbon intensity of clients	Average carbon intensity of energy grid used by clients	Location of clients (IP)	Float [20,795]	$(795 - output)/(795 - 20)$
Carbon intensity server	Carbon intensity of energy grid used by the server	Location of server (IP)	Float [20,795]	$(795 - output)/(795 - 20)$
		Notion: Hardware Efficiency		
Avg. hardware efficiency of clients	Average performance per watt (CPU or GPU Mark/ TDP) of CPUs and GPUs used by clients	CPU and GPU models of clients	Float [20,1447]	$(1447 - output)/(1447 - 20)$
Hardware efficiency of clients	Performance per watt (CPU or GPU Mark/ TDP) of CPUs and GPUs used by the server	CPU and GPU models of server	Float [20,1447]	$(1447 - output)/(1447 - 20)$
		Notion: Federation Complexity		
Number of global training rounds	Number of global training rounds in the FL system	Config file	Integer	output:[10, 10^2, 10^3, 10^4, 10^5, 10^6] norm:[1, 0.8, 0.6, 0.4, 0.2, 0]
Number of clients	Number of clients in the federation	Config file	Integer	output:[10, 10^2, 10^3, 10^4, 10^5, 10^6] norm:[1, 0.8, 0.6, 0.4, 0.2, 0]
Client selection rate	Percentage of clients selected in each training round to share their models	Config file	Float [0,1]	[0,1]
Average number of local training rounds	Average number of local training rounds performed by clients within one global training round	Config file	Integer	output:[10, 10^2, 10^3, 10^4, 10^5, 10^6] norm:[1, 0.8, 0.6, 0.4, 0.2, 0]
Average dataset size	Average number of samples used by clients in one training round	Client Statistics	Integer	output:[10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}] norm:[1, 0.8, 0.6, 0.4, 0.2, 0]
Model size	Number of features/depth of decision tree/number of parameters in NN	Model	Integer	output:[10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}] norm:[1, 0.8, 0.6, 0.4, 0.2, 0]

- PC_1: The algorithm must not store sensitive data from the FL model.
- PC_2: The algorithm must not leak or share sensitive data from clients, the server, and the FL model with third parties. Additionally, to compute the carbon emission pillar no sensitive data used to train local models must be shared.
- PC_3: The metrics calculations can occur at the client's local devices, the central server, or collaboratively between both.

4.1. Sustainability Pillar: Notions and Metrics

Table 2 shows the notions and metrics explained in Section 3 and considered in the algorithm for the sustainability pillar. Descriptions, inputs, outputs, and normalization details are provided for each metric. For metric computation, the CodeCarbon package [4] is leveraged to obtain the emissions related to the hardware employed by the server/clients and the emissions related to the location of the nodes in the FL setup. This package has been selected by the most representative solutions in the literature, as described in Section 2. Besides, for the calculation of *Hardware Efficiency metrics*, the most popular benchmarking software for processors is PassMark [16]. It computes a performance score by running standardized tests that simulate real-world workloads, such as executing complex mathematical calculations. PassMark has provided a database with Power Performance measurement for over 3000 CPUs and 2000 GPUs published on Kaggle, which can be used to evaluate the client and server processor efficiency in the algorithmic prototype design.

In addition to the previous ten metrics, the proposed algorithm also implements the 41 metrics belonging to the remaining six pillars proposed in [20].

4.2. Algorithm Design

Figure 2 shows the overview of the algorithm design. The proposed algorithm considers the following inputs.

- *Emissions*. It contains the IP of clients and server, CPU and GPU models, and config files of the federation needed to compute the ten sustainability metrics (see Table 2).
- *FL Model*. It contains information about the model configuration and model personalization.
- *FL Framework Configuration*. It contains information about the number of clients, the client selection mechanisms, the aggregation algorithm, and the model hyperparameters.
- *FactSheet*. It contains essential details for the accountability of the training process, federation, and the individuals involved [1].
- *Statistics*. It contains information about the client class balance, client test performance loss, client test accuracy, client clever score, client feature importance, client participation rate, client class imbalance, client average training time, model size, and average upload-/download bytes.

These input sources serve as the foundation for deriving the sustainability metrics outlined in Table 2 and the metrics belonging to the remaining six pillars proposed in [20]. The resulting metric values are then normalized to ensure a consistent range. It is essential to note that each metric can encompass distinct input sources and may be computed at different stages of the federated learning (FL) model creation process, namely pre-training, during-training, or

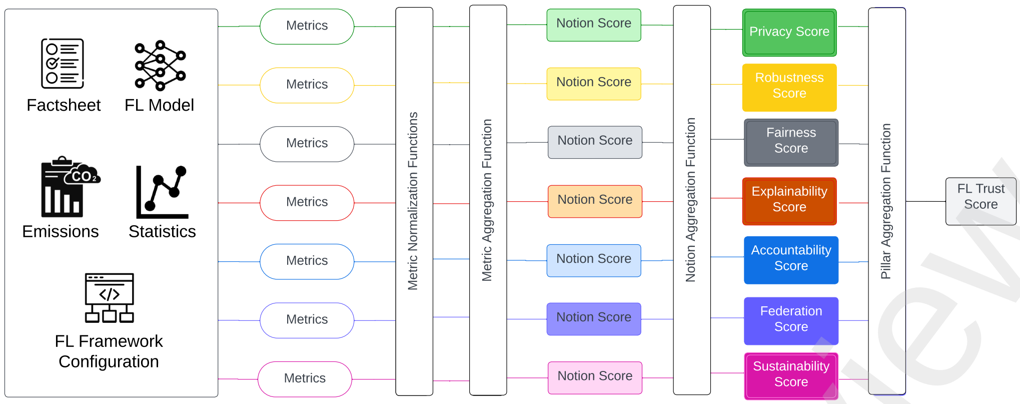

Figure 2: Algorithm Design

post-training, by various participants within the federation, be it clients or servers. Once the normalized metric outputs are determined, they are assigned adaptive weights (defined by end-users according to the scenario particularities) and combined to produce a score for each notion. Each pillar incorporates one or more notions, assessed based on predefined yet adjustable weights for each metric. Consequently, the same procedure is reiterated to derive pillar scores through the weighting and aggregation of notion scores. Ultimately, the overall trust score of the FL model is determined as a custom amalgamation of the pillar scores.

4.3. Algorithm Deployment

Once designed, the algorithm was implemented and deployed in a well-known FL framework called FederatedScope [26]. After the deployment, the following steps show how the sustainability and trustworthy FL scores are calculated.

1. *Setup*: Start the federation by initiating FederatedScope. It takes the federation configuration as input and initiates clients and the server, as well as the proposed trustworthiness calculation algorithm. Additionally, it populates the FactSheet with pre-training metrics such as the number of clients in the federation and the number of training rounds.
2. *Model Broadcast*: The server broadcasts the global model to selected clients in the federation.
3. *Local Training*: The selected clients train their local models with their local private dataset. At this point, clients use the CodeCarbon package to obtain metrics relevant to the sustainability pillar computation.
4. *Report Emissions Metrics*: Selected clients report metrics such as the hardware models and energy grid, which are stored in the Emissions file.
5. *Model Sharing*: Selected clients then share their updated model parameters with the server.
6. *Federated Aggregation*: the Aggregator is used by the server and performs secure aggregation over the model updates received from selected clients.
7. *Evaluation*: After each training round, the clients perform model evaluation and call the proposed algorithm to perform metric calculations.
8. *Next training round*: Steps two to eight are repeated until all the training rounds are finished.
9. *Propagate Evaluation Results*: Once the final training round is finished and the collaborative training stops, the evaluation results get propagated to the FactSheet through the algorithm.
10. *Trust Score Computation*: The algorithm computes the overall trust score from the FactSheet and report, including the trustworthiness scores stored in the output directory of FederatedTrust.

The execution of the FederatedScope training process, together with the evaluation of the FL sustainability and trustworthiness, is depicted in Algorithm 1.

5. Evaluation and Results

This section evaluates the proposed algorithm through a pool of experiments. Firstly, it includes a quantitative analysis of its functionality. Then, it analyzes how the proposed system can effectively help users to better understand the sustainability of the FL systems and support decision-making processes.

5.1. Functionality Evaluation

Four use cases (UC) are conducted to examine the functionality of the sustainability pillar. They consider several levels of federation complexity, diverse degrees of carbon intensity in the energy grid utilized by both clients and the server, and different hardware efficiencies of the CPUs employed by the clients and the server. The setups for these four cases are depicted in Table 3. In the following experiments,

Algorithm 1 Training in FederatedScope

Input: N clients, sampling size m, central server S, total number of iterations T, initial model $\overline{w}(0)$, setup configurations C, FederatedTrust metric manager ft

Output: Evaluation results, trustworthiness report, estimated carbon emissions

```
S sends the hashed ids of all clients i ∈ [N] and C to ft
ft creates FactSheet with information from C
ft creates a map of hashed client IDs to values of 0, representing the initial selection rate
S sends the model metadata to ft
S requests class distribution information from all clients i ∈ [N]
ft creates emissions file ef
for each client i ∈ [N] do
    Client i uses ft function to calculate the sample size per class of local data
    ft creates or updates the class distribution map of hashed labels to sample size
end for
for t = 0 to T do
    S randomly samples D(t) ⊂ [N] clients with size of m
    S sends the hashed IDs of the selected clients to ft
    ft updates the client selection rate map
    S broadcasts the current model w̄(t) to all clients i ∈ D(t)
    for each client i ∈ D(t) do
        Client i initializes an EmissionsTracker object from CodeCarbon and starts emissions tracking for training
        Client i performs local training with w̄(t)
        Client i uses ft function to stop emission tracking for training and to save results
        ft updates ef
        Client i sends new model updates w(t + 1)_i back to S
    end for
    S initializes an EmissionsTracker object from CodeCarbon and starts emissions tracking for aggregation
    S performs secure aggregation of all updates received into a new global model w̄(t + 1)
    S uses ft function to stop emissions tracking for aggregation and to save results
    ft updates ef
end for
S sends final global model w̄' to every client i ∈ [N] for performance evaluation
for each client i ∈ [N] do
    Client i computes evaluation metrics with local test data and global model w̄'
    Client i sends the evaluation results back to S
end for
S aggregates the evaluation results and sends them to ft
ft receives the evaluation results and populates the FactSheet with them
S asks ft to evaluate the trustworthiness of the model
ft computes the trustworthiness score and estimated emissions and generates a report JSON and print message
```

each metric carries equal weight when calculating the notion score. In addition, when determining the sustainability pillar score, the carbon intensity of the energy source metric is assigned a weight of 0.5, while the hardware efficiency and federation complexity metrics are each assigned a weight of 0.25.

Table 3
Setups for Functionality Evaluation Experiment

	UC A	*UC B*	*UC C*	*UC D*
Clients Loc.	Albania	50% in Kosovo 50% in Gambia	Switzerland	South Africa
Server Loc.	Albania	South Africa	Switzerland	South Africa
Clients Hardware	i7-1250U	AMD FX-9590	40% E5-4620 35% E5-4627 25% E5-2650	i5-1335U
Server Hardware	i7-1250U	W2104	E5-4620	i7-1250U
Rounds	10	1000	1000	10
No. of Clients	5	1000	1000	8
Selection Rate	0.2	1	0.8	0.3
Local Rounds	1	90	90	1
Dataset Size	100	1.10E+06	1.10E+06	100
Model size	98,000	1.00E+13	1.00E+13	99,300

5.1.1. Low Carbon Intensity and High Hardware Efficiency

UC A represents the optimal situation with minimal CO_2eq emissions. In this scenario, the server and all five clients utilize the Intel Core i7-1250U CPU, which boasts exceptional efficiency with a power performance of 1447, the greatest recorded by PassMark thus far. Moreover, the federation complexity remains low, characterized by a limited number of clients, global and local training rounds, as well as a small client selection rate, dataset size, and model size. Furthermore, both the clients and server are situated in Albania, which possesses one of the least carbon-intensive energy grids. Therefore, as depicted in Table 4, UC A obtains a carbon intensity of energy source notion score of 1, a hardware efficiency notion score of 1, and a federation complexity notion score of 0.98, resulting in the highest result with an overall sustainability score.

5.1.2. High Carbon Intensity and Low Hardware Efficiency

UC B illustrates a worst-case scenario with inefficient hardware, a highly complex federation resulting in high energy consumption, and high carbon intensity of the electricity grid used, resulting in substantial CO_2eq emissions. In this scenario, a server utilizes an Intel Xeon W-2104 CPU with a low power performance measurement of 51.67. All 1000 clients have an AMD FX-9590 CPU, which exhibits a low power performance of 30.76. Consequently, the overall hardware used shows inefficiency, achieving a hardware efficiency notion score of 0.01. Moreover, the federation complexity is significant, involving 1000 global and 90 local training rounds, with a federation complexity notion score of 0.13. Additionally, the server is located in South Africa, which has one of the most energy-intensive grids, emitting 709g CO_2eq per kWh. Half of the clients are situated in Kosovo, which operates a carbon-intensive energy grid generating 769g of CO_2eq per kWh. The other half of the clients are based in Gambia, which relies on a carbon-intensive electricity grid releasing 700g of CO_2eq per kWh. Consequently, the average carbon intensity of the electricity

Table 4
Sustainability Score for Functionality Evaluation

Metric	*UC A*	*UC B*	*UC C*	*UC D*
Sustainability Pillar	1.00	0.09	0.55	0.53
- Carbon Intensity of Energy Source Notion (weight 0.5)	1.00	0.09	1.00	0.11
- - Avg. Carbon Intensity of Energy Grid Clients	1.00	0.08	1.00	0.11
- - Carbon Intensity of Energy Grid Server	1.00	0.11	1.00	0.11
- Hardware Efficiency Notion (weight 0.25)	1.00	0.01	0.04	0.94
- - Avg. Hardware Efficiency Clients	1.00	0.01	0.05	0.87
- - Hardware Efficiency Server	1.00	0.02	0.04	1.00
- Federation Complexity Notion (weight 0.25)	0.98	0.13	0.17	0.96
- - Number of Training Rounds	1.00	0.17	0.17	1.00
- - Number of Clients	1.00	0.17	0.17	1.00
- - Client Selection Rate	0.89	0.00	0.22	0.77
- - Avg. Number of Local Training Rounds	1.00	0.17	0.10	1.00
- - Average Dataset Size	1.00	0.20	0.20	1.00
- - Model Size	1.00	0.14	0.14	1.00

grid used by clients totals 734.5g of CO_2eq per kWh, with a carbon intensity of energy source notion score of 0.09. Combining these three notions with the weighted average, the overall score for the sustainability pillar is 0.09 for UC B, which represents a worst-case scenario in terms of the sustainability pillar using inefficient hardware and carbon-intensive electricity grids in combination with a complex federation.

5.1.3. *Low Carbon Intensity and Low Hardware Efficiency*

UC C represents a scenario where the hardware used is inefficient, and the federation is complex, leading to high energy consumption. However, the carbon intensity of the electricity grid is low, resulting in medium CO_2eq emissions. In this case, the server utilizes an Intel Core i7-6800K CPU with a power performance of 76.29. Among the clients, 40% use an Intel Xeon E5-4620 CPU with a power performance of 100.24, 35% use an Intel Xeon E5-4627 with a power performance of 71.69, and 25% use an Intel Xeon E5-2650 with a power performance of 105.21. Overall, the hardware is considered inefficient, achieving a hardware efficiency notion score of 0.01. The federation complexity is high, with a large number of clients, global training rounds, local training rounds, and parameters in the DNN model, resulting in a federation complexity notion score of 0.17. However, both the server and clients are located in Switzerland, where the energy grid has a low carbon intensity of 32g CO_2eq per kWh, achieving a carbon intensity of energy source notion score of 1. By combining these three notions, the overall score for the sustainability pillar is 0.55 for UC C.

5.1.4. *High Carbon Intensity and High Hardware Efficiency*

UC D utilizes highly efficient computational hardware but has a high carbon intensity in its grid, leading to a moderate level of CO_2eq emissions, in contrast to UC C. In UC D, the server utilizes the Intel Core i7-1250U CPU power performance of 1447, while all eight clients use the Intel Core i5-1335U with a power performance of 1268. Additionally, the federation complexity is low, with a small number of clients, global training rounds, local training rounds, and a small client selection rate, dataset size, and model size. Consequently, the hardware efficiency notion score and federation complexity notion score are 0.94 and 0.96, respectively. However, both the clients and server are situated in South Africa, where the carbon intensity of the energy source is 709g CO_2eq per kWh, resulting in a carbon intensity of energy source notion score of 0.11. Therefore, the final sustainability score is 0.53, similar to UC C.

In conclusion, this experiment provides empirical evidence supporting the effectiveness of the proposed sustainability pillar, which enables the quantitative measurement of CO_2eq emissions in various contexts of FL systems. Moreover, this sustainability pillar yields accurate and interpretable sustainability scores based on such measurements.

5.2. Effectiveness Evaluation

Nevertheless, validating the calculated sustainability pillar could enhance the credibility of the trust score is a complex task. This difficulty primarily stems from the absence of the ground truth, rendering quantitative analysis notably challenging. Therefore, this experiment analyzes and validates the effectiveness and value-adding properties of the sustainability pillar through a hypothetical case study.

Assuming a multinational IT consulting company based in Luxembourg, with two research and development centers located in Zurich, Switzerland, and Johannesburg, South Africa. Both branches have simultaneously proposed an FL-based training proposal, with their respective training configurations outlined in the Table 5. However, due to limited resources, only one proposal can be implemented. As the director of the research and development centers, the decision-maker aims to follow the guidance of the AI-HLEG. It intends to evaluate the trust score of the two proposals using the algorithm proposed in this work. This calculation will ultimately determine which proposal should be adopted.

Table 5 presents the configurations of the two proposals, which exhibit a high degree of similarity. The primary distinction lies in the fact that Proposal A, involving the Johannesburg team, necessitates a greater number of clients to participate in the training process and entails a substantially higher number of training rounds compared to Proposal B, which is proposed by the Zurich team. Additionally, both teams intend to conduct the training process at their local facilities.

The director utilized the proposed system to upload the proposals submitted by the two teams. This system then computed and evaluated the scores of various pillars, such as

Table 5
The FL Configuration of the Proposals from the Two Branches

Metric	*Proposal A*	*Proposal B*
Model	ConvNet2	ConvNet2
Local Rounds	100	10
Dataset	FEMNIST	FEMNIST
Data Split (Train, Val., Test)	0.6/0.2/0.2	0.6/0.2/0.2
Batch Size	50	50
Loss	CrossEntropyLoss	CrossEntropyLoss
Consistent Label Distribution	False	False
Number of Clients	1000	10
Client Selection Rate	0.3	0.6
Federation Rounds	1000	10
Clients Hardware	Intel i7-8650U	Intel i7-8650U
Server Hardware	Intel i7-8650U	Intel i7-8650U
Client Location	South Africa	Switzerland
Server Location	South Africa	Switzerland
Differential Privacy	Epsilon 10	Epsilon 10
Aggregation Method	FedAvg	FedAvg

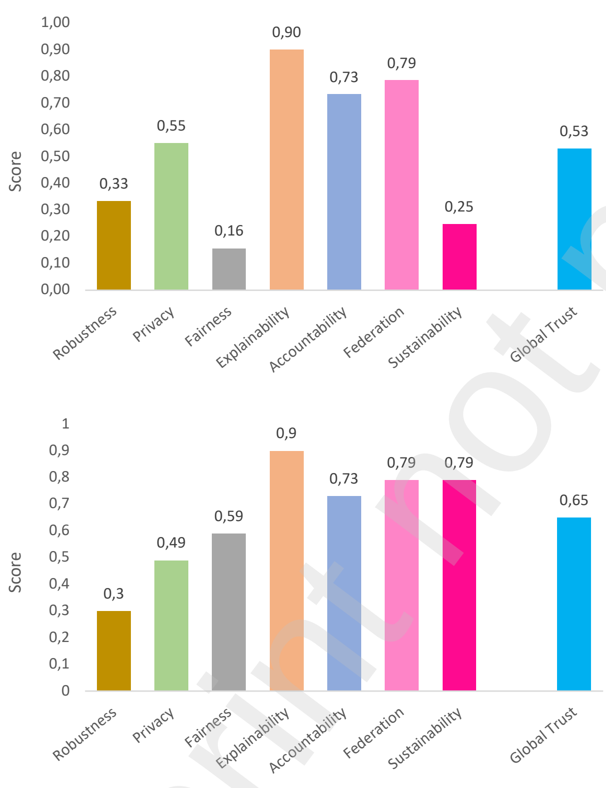

Figure 3: Results of Evaluation of the Proposed Algorithm for Proposal A (top) and Proposal B (bottom)

robustness, privacy, and fairness, ultimately aggregating them to generate a trust score. In this experiment, equal weight was assigned to all the pillars during calculations.

However, there is a lack of established methods, equations, and practical calculation techniques for computing all the notions and metrics mentioned in Section 3. As a result, the goal of the implementation is to create a simplified prototype that incorporates basic principles, concepts, and metrics that can be calculated. All the computed notions are presented in Table 6.

Table 6
Pillar and Notion Scores for Two Proposals

Metric	*Proposal A*	*Proposal B*
Robustness Pillar	0.33	0.30
- Resilience to Attacks	0.27	0.40
- Algorithmic Robustness	0.51	0.00
- Client Reliability	0.23	0.50
Privacy Pillar	0.55	0.49
- Differential Privacy	1.00	1.00
- Indistinguishability	0.00	0.00
- Uncertainty	0.65	0.47
Fairness Pillar	0.16	0.59
- Selection Fairness	0.47	0.76
- Performance Fairness	0.00	1.00
- Class Distribution	0.00	0.00
Explainability Pillar	0.90	0.90
- Interpreability	0.80	0.80
- Post-hoc Methods	1.00	1.00
Accountability Pillar	0.73	0.73
- Factsheet Completeness	0.73	0.73
Federation Pillar	0.79	0.79
- Client Management	1.00	1.00
- Optimization	0.57	0.57
Sustainability Pillar	0.25	0.79
- Carbon Intensity of Energy Grid Server	0.11	0.98
- Hardware Efficiency	0.28	0.28
- Federation Complexity	0.49	0.91

The results of the system, as depicted in Figure 3, indicate that both proposals have similar scores in different aspects, including explainability, accountability, and federation. This similarity can be attributed to the proximity of their respective configurations. As indicated in the Table 6, both proposals demonstrated low levels of robustness as they were not optimized for resisting attacks. Regarding privacy, proposal B outperformed proposal A due to its significant number of nodes, which introduced more uncertainty and improved overall privacy. Besides, proposal B exhibited a greater fairness score compared to proposal A due to its superior level of client selection fairness, and the performance of the model among the clients is even.

Before the inclusion of the sustainability pillar, the trust scores for the two proposals were relatively similar, with proposal A receiving a score of 0.58 and proposal B receiving a score of 0.63, indicating a minimal difference of 0.05. This posed a challenge in determining which proposal aligned more closely with the concept of trustworthiness. However, with the introduction of the sustainability pillar, the data presented in the Table 6 reveals that proposal B exhibited notable advantages regarding carbon intensity of energy source and federation complexity. As a result, the final trust scores were adjusted to 0.53 and 0.65 for proposal A and proposal B, respectively, resulting in an increased discrepancy of 0.12. Ultimately, proposal B emerged as the winner due to its superior performance in sustainability.

In summary, this experiment serves as a hypothetical case study to illustrate that the sustainability pillar effectively enhances users' comprehension of the environmental impacts of using FL systems and offers valuable assistance in the decision-making process.

6. Discussion

This section discusses the most relevant limitations noticed during the design and implementation process of the proposed algorithm. The intention is to seek future improvements and iterations over the pillar notions, metrics, and their calculation process.

Coming to limitations in terms of the sustainability pillar, the magnitude in which the single metrics influence the CO_2eq emissions are uncertain but are weighted equally. For example, the number of training rounds and clients in the federation have the same weight in this prototype design. Still, more training rounds might contribute more to the final CO_2eq emissions than the number of clients in the federation. Similarly, at the notion level, it is unclear if the efficiency of the hardware notion and the federation complexity notion influence the CO_2eq emissions equally. Thus, the weighting of the metrics and notions might only partially reflect their influence on the federation's environmental impact. So, weighs can be established according to the the scenario particularities and further investigation is needed in order to define proper weighs per pillar and metric.

For the carbon intensity of the energy source notion, the average carbon intensity of the country's energy grid is an approximation. It is due to the carbon intensity of the electricity grid fluctuates within a country and a day or season. However, for the purpose used, it is a fairly good approximation. Regarding the hardware efficiency notion, only the efficiency of CPUs and GPUs is considered. To be more accurate, the efficiency of other components, such as RAM, could be integrated. Additionally, the power performance metric depends on PassMark benchmarking scores and is not comparable to other benchmarking software scores. Further, if emissions want to be measured, the emitted CO_2eq for producing the hardware should be included. This, however, is fairly difficult to do.

Finally, there are some additional aspects, like privacy-preserving technologies used in the federation, that might be relevant for carbon emissions estimations. For example, if a federation uses homomorphic encryption as privacy protection, it would increase energy consumption and emissions due to its computational complexity. Further, FL systems often use methods to detect malicious clients or free-riders, such as clustering or the H-MINE algorithm. Such methodologies are computationally heavy and may increase the computational costs, energy consumption, and CO_2eq emissions. Finally, despite data used to train local models is not shared to compute carbon emissions, further investigation is needed to assess if sensitive data could be inferred by curious members of the federation or attackers.

7. Conclusion and Future Work

This work introduces the sustainability pillar to the trustworthy FL taxonomy, aiming to assess the environmental impact of FL systems. This new pillar comprises ten metrics belonging to three main notions: hardware efficiency, federation complexity, and the carbon intensity of the energy grid. Together, these notions provide a comprehensive evaluation of an FL system's resource consumption and environmental impact, highlighting the importance of efficient hardware and low-carbon energy sources. Additionally, this work designs and implements an algorithm for evaluating FL trustworthiness by incorporating the sustainability pillar. Using the CodeCarbon Python package, the algorithm now considers the hardware models used and the carbon intensity of the energy grid based on the geographical locations of clients and servers. Extensive evaluations across various scenarios reveal that FL systems with low complexity, efficient hardware, and a clean energy grid receive high sustainability and trustworthiness scores.

Future work will refine the sustainability scores by investigating and adjusting the weights of individual metrics related to carbon emissions and other pillars. This includes considering the computational costs of privacy-preserving methods, like Differential Privacy and Homomorphic Encryption, and malicious client detection techniques. Enhancing the security of the FederatedTrust prototype, expanding its compatibility with various frameworks, and adapting it to decentralized federations are also potential avenues for improvement. Additionally, incorporating unimplemented metrics from the other six pillars of the taxonomy could further enhance the prototype's comprehensiveness.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Chao Feng. Methodology, Conceptualization, Writing, Review & Editing. **Alberto Huertas Celdrán.** Methodology, Writing, Review. **Pedro Miguel Sánchez Sánchez.** Methodology, Writing, Review. **Lynn Zumtaugwald.** Data curation. **Gérôme Bovet.** Project administration, Funding acquisition. **Burkhard Stiller.** Supervision, Funding acquisition.

Acknowledgment

This work has been partially supported by *(a)* the Swiss Federal Office for Defense Procurement (armasuisse) with the CyberMind and DATRIS (CYD-C-2020003) projects, *(b)* the University of Zürich UZH, and *(c)* the strategic project DEFENDER from the Spanish National Institute of Cybersecurity (INCIBE) by the Recovery, Transformation, and Resilience Plan, Next Generation EU.

References

[1] Arnold, M., Bellamy, R.K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K.N., Olteanu, A., Piorkowski, D., et al., 2019. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. IBM Journal of Research and Development 63, 6–1.

[2] Beltrán, E.T.M., Pérez, M.Q., Sánchez, P.M.S., Bernal, S.L., Bovet, G., Pérez, M.G., Pérez, G.M., Celdrán, A.H., 2023. Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. IEEE Communications Surveys & Tutorials , 1–1.

[3] Celdran, A.H., Kreischer, J., Demirci, M., Leupp, J., Sanchez, P.M., Franco, M.F., Bovet, G., Perez, G.M., Stiller, B., 2023. A framework quantifying trustworthiness of supervised machine and deep learning models, in: SafeAI2023: The AAAI's Workshop on Artificial Intelligence Safety, pp. 2938–2948.

[4] CodeCarbon, . Codecarbon. https://codecarbon.io. Accessed: 22.02.2023.

[5] George, A.S., George, A.H., Martin, A.G., 2023. The environmental impact of ai: A case study of water consumption by chat gpt. Partners Universal International Innovation Journal 1, 97–104.

[6] He, M., Li, Z., Liu, C., Shi, D., Tan, Z., 2020. Deployment of artificial intelligence in real-world practice: opportunity and challenge. The Asia-Pacific Journal of Ophthalmology 9, 299–307.

[7] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., Pineau, J., 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. arXiv:2002.05651.

[8] HLEG, A., . Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html. Accessed: 15.02.2023.

[9] Kaur, D., Uslu, S., Rittichier, K.J., Durresi, A., 2022. Trustworthy artificial intelligence: a review. ACM Computing Surveys (CSUR) 55, 1–38.

[10] Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T., 2019. Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700 .

[11] Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., Tang, J., 2022. Trustworthy ai: A computational perspective. ACM Transactions on Intelligent Systems and Technology 14, 1–59.

[12] Luccioni, A.S., Hernandez-Garcia, A., 2023. Counting carbon: A survey of factors influencing the emissions of machine learning. arXiv preprint arXiv:2302.08476 .

[13] Luccioni, A.S., Viguier, S., Ligozat, A.L., 2022. Estimating the carbon footprint of bloom, a 176b parameter language model. arXiv preprint arXiv:2211.02001 .

[14] Martonosi, M., Brooks, D., Bose, P., 2001. Modeling and analyzing cpu power and performance: Metrics, methods, and abstractions. SIGMETRICS 2001/Performance 2001-Tutorials .

[15] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR. pp. 1273–1282.

[16] PassMark, 2023. Thermal Design Power. https://www.cpubenchmark.net/power_performance.html. Accessed: 03.05.2023.

[17] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.M., Rothchild, D., So, D., Texier, M., Dean, J., 2021. Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350 .

[18] Petroleum, B., 2022. Statistical review of world energy 2022. bp.

[19] Qiu, X., Parcollet, T., Fernandez-Marques, J., de Gusmao, P.P., Gao, Y., Beutel, D.J., Topal, T., Mathur, A., Lane, N.D., 2023. A first look into the carbon footprint of federated learning. J. Mach. Learn. Res. 24, 129–1.

[20] Sánchez, P.M.S., Celdrán, A.H., Xie, N., Bovet, G., Pérez, G.M., Stiller, B., In Press. Federatedtrust: A solution for trustworthy federated learning. Future Generations Computer Systems .

[21] Schlömer, S., Bruckner, T., Fulton, L., Hertwich, E., McKinnon, A., Perczyk, D., Roy, J., Schaeffer, R., Sims, R., Smith, P., et al., 2014. Annex iii: Technology-specific cost and performance parameters, in: Climate change 2014: Mitigation of climate change: Contribution of working group III to the fifth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press, pp. 1329–1356.

[22] Shi, Y., Yu, H., Leung, C., 2021. A survey of fairness-aware federated learning. arXiv preprint arXiv:2111.01872 .

[23] Smuha, N.A., 2019. The eu approach to ethics guidelines for trustworthy artificial intelligence. Computer Law Review International 20, 97–106.

[24] Strubell, E., Ganesh, A., McCallum, A., 2019. Energy and policy considerations for deep learning in nlp. arXiv preprint arXiv:1906.02243 .

[25] Tariq, A., Serhani, M.A., Sallabi, F., Qayyum, T., Barka, E.S., Shuaib, K.A., 2023. Trustworthy federated learning: A survey. arXiv preprint arXiv:2305.11537 .

[26] Xie, Y., Wang, Z., Gao, D., Chen, D., Yao, L., Kuang, W., Li, Y., Ding, B., Zhou, J., 2022. Federatedscope: A flexible federated learning platform for heterogeneity. arXiv preprint arXiv:2204.05011 .

[27] Zhang, Y., Zeng, D., Luo, J., Xu, Z., King, I., 2023. A survey of trustworthy federated learning with perspectives on security, robustness, and privacy. arXiv preprint arXiv:2302.10637 .

[28] Zumtaugwald, L., 2023. Algorithm to Compute the Sustainability and Trustworthiness of FL. https://github.com/lzumta/FederatedScope. Last Visit Oct. 2023.