Assessing the Sustainability and Trustworthiness of Federated Learning Models

Chao Feng[1], Alberto Huertas Celdrán[1,2], Pedro Miguel Sánchez Sánchez[2], Lynn Zumtaugwald[1],
Gérôme Bovet[3], Burkhard Stiller[1]
[1]Communication Systems Group, Department of Informatics, University of Zurich UZH, CH–8050 Zürich, Switzerland
[cfeng, huertas, stiller]@ifi.uzh.ch, lynn.zumtaugwald@uzh.ch
[2]Department of Information and Communications Engineering, University of Murcia, 30100 Murcia, Spain
pedromiguel.sanchez@um.es
[3]Cyber-Defence Campus, armasuisse Science & Technology, CH–3602 Thun, Switzerland
gerome.bovet@armasuisse.ch

***Abstract*—Artificial intelligence (AI) increasingly influences critical decision-making across sectors. Federated Learning (FL), as a privacy-preserving collaborative AI paradigm, not only enhances data protection but also holds significant promise for intelligent network management, including distributed monitoring, adaptive control, and edge intelligence. Although the trustworthiness of FL systems has received growing attention, the sustainability dimension remains insufficiently explored, despite its importance for scalable real-world deployment. To address this gap, this work introduces sustainability as a distinct pillar within a comprehensive trustworthy FL taxonomy, consistent with AI-HLEG guidelines. This pillar includes three key aspects: hardware efficiency, federation complexity, and the carbon intensity of energy sources. Experiments using the FederatedScope framework under diverse scenarios, including varying participants, system complexity, hardware, and energy configurations, validate the practicality of the approach. Results show that incorporating sustainability into FL evaluation supports environmentally responsible deployment, enabling more efficient, adaptive, and trustworthy network services and management AI models.**

***Index Terms*—Sustainable AI, Carbon Footprint, Federated Learning.**

I. Introduction

As Artificial Intelligence (AI) systems become increasingly embedded in critical infrastructures, including communication networks and cloud-edge services, their role in intelligent network management, such as dynamic resource allocation, traffic optimization, and fault detection, has grown significantly [1]. Deep Learning (DL), as the backbone of many AI systems, entails intensive computational demands during training and inference, resulting in substantial carbon emissions. In large-scale networked environments, such as edge-cloud infrastructures and distributed systems, this resource consumption increases with the number of participating agents, posing a growing sustainability concern. Moreover, DL's dependence on massive datasets often leads to inefficient data handling, adding further energy overhead. These challenges are particularly prominent in network management tasks, where AI is employed for dynamic resource allocation, traffic control, and fault diagnosis across heterogeneous infrastructures.

Sustainability in AI extends beyond energy and computation. Ethical concerns, such as bias, privacy violations, and discrimination, are especially critical in decentralized and privacy-sensitive domains like personalized network services. Thus, sustainable AI must be pursued in conjunction with broader pillars of trustworthiness, including robustness, transparency, fairness, and accountability.

To this end, ensuring AI trustworthiness has become a global priority. Regulatory initiatives, such as those led by the European Commission's High-Level Expert Group on AI (AI-HLEG) [2], have established comprehensive guidelines. The AI-HLEG outlines seven core requirements for trustworthy AI: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, fairness, environmental well-being, and accountability [3].

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) [4], an innovative paradigm that enables multiple clients to collaboratively train models without necessitating the exchange of private data. Today, FL confronts multifaceted challenges, spanning scalability, single point of failure, architectural design, or privacy and security concerns, among others. However, while FL inherently incorporates privacy-preserving features, trustworthy AI remains a pivotal dimension within FL systems [5].

In this context, prior work [10], [8] defined a baseline by formulating taxonomies for trustworthy ML, DL, and FL. Other work, such as [5], implemented algorithms and frameworks for assessing the trustworthiness of FL systems. However, environmental well-being is missing in full there. Specifically, key sustainability factors such as CO_2eq emissions, hardware efficiency, federation complexity, and energy grid carbon intensity, highlighted by AI-HLEG, remain unaddressed, despite their importance for optimizing FL configurations and promoting awareness. Moreover, unlike centralized paradigms, FL operates in a decentralized manner, involving heterogeneous and geographically distributed clients. This results in substantial variability in computational capabilities and local carbon intensities across nodes, making sustainability assessment more complex. FL's dynamic and distributed nature increases the challenge of designing effective, context-aware environmental impact evaluation metrics.

978-3-903176-75-1 ©2025 IFIP

TABLE I: 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. [6] (2021)	No	Yes	No	No	No	Partially	No
Liu et al. [7] (2022)	Yes	No	Yes	No	No	Partially	No
Tariq et al. [8] (2023)	Yes	Yes	Yes	No	Yes	No	No
Zhang et al. [9] (2023)	Yes	No	Yes	No	No	No	No
Sanchez et al.[5] (2023)	Yes	Yes	Yes	Yes	Yes	Yes	No
This work	Yes	Yes	Yes	Yes	Yes	Yes	Yes

This work makes three key contributions toward sustainable and trustworthy Federated Learning (FL). First, it proposes a unified taxonomy for Trustworthy FL with seven pillars: privacy, robustness, fairness, accountability, federation, explainability, and a newly introduced sustainability pillar, which encompasses carbon intensity, hardware efficiency, and federation complexity, along with ten corresponding metrics. Second, it presents an evaluation algorithm (available in [11]) that incorporates sustainability metrics into the trustworthiness assessment. Third, the algorithm is integrated into the FederatedScope framework and validated across diverse settings, demonstrating its effectiveness in evaluating FL trustworthiness. By explicitly addressing sustainability, the proposed approach supports more efficient and responsible FL-based network management approaches, including intelligent orchestration, edge coordination, and energy-aware optimization.

II. Related Work

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

A. Trustworthy FL Evaluation

TABLE I summarizes the existing trustworthy FL taxonomies and their coverage of trustworthy FL pillars defined by the AI-HLEG. The taxonomy from Shi et al. [6] 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. [7] provided a taxonomy covering the pillars of privacy, robustness, and partially the pillar of federation. Tariq et al. [8] 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. [9] 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. [5]. 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 FL.

After reviewing the literature, the most important limitation becomes apparent 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 [2]. Since [5] 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.

B. Estimating Emissions of AI/FL

Most works focus on estimating the carbon emissions of specific models. Lucconi et al. [12] provided a survey on aspects that influence the CO_2eq of ML. Luccioni et al. Patterson et al. [13] 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. In the field of FL, Qui et al. [14] provided a first look into the carbon footprint of FL models by incorporating parameters specific to FL and comparing the emissions produced by FL models with those produced by centralized ML models. Similarly to estimating the carbon emissions of AI/FL models, tools have been developed to track carbon emissions. CodeCarbon [15] and the Experimental Emissions Tracker [16] can be used to track emissions during the training process, while the ML CO_2eq Calculator [17] can be used to calculate the emissions after training.

Despite the effort and work done in this research field, none have 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 [2].

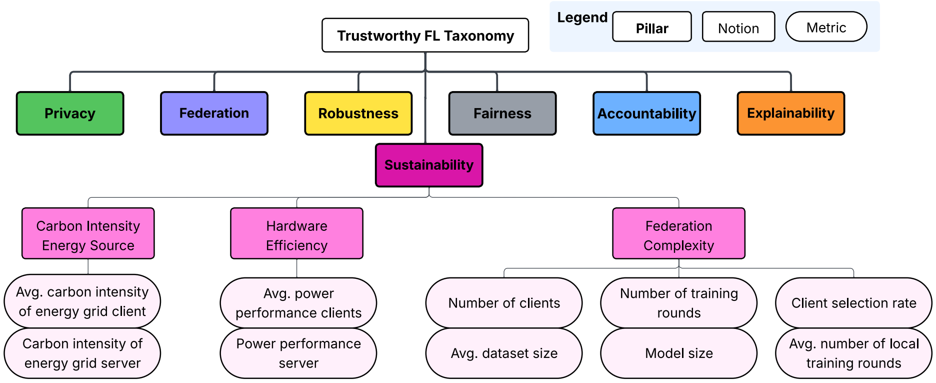

Fig. 1: Trustworthy FL Taxonomy

III. The Sustainability Pillar of Trustworthy FL

The trustworthy FL taxonomy is structured hierarchically into three levels. At the top level, **pillars** represent the fundamental aspects to be considered, such as privacy, robustness, and sustainability. Each pillar is decomposed into several **notions**, which capture the specific dimensions that need to be addressed to achieve the corresponding pillar. Finally, each notion is associated with one or more **metrics** that allow its quantitative assessment.

This section describes the notions and metrics that comprise trustworthy FL's sustainability pillar. This pillar includes the energy grid's carbon intensity, the underlying hardware's efficiency, and the federation's complexity. The overview of the Trustworthy FL framework is presented in Fig. 1.

A. Carbon Intensity

Carbon intensity depends on the energy source used for electricity generation. The energy mix used to train FL models significantly impacts total emissions. For example, training with 500 kWh from coal results in 410 kg CO_2eq, versus just 5.5 kg with nuclear power. National grids vary widely, ranging from 20g in Lesotho to 795g CO_2eq/kWh in Botswana [18]. 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 [19]. 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.

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

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

B. 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 [17]. 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) [20]. It is important to note that none of these metrics provides 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 [21]. 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 [21]. It is measured in performance per Watt using Equation 2 and 3.

$$H_E = \frac{H_{BP}}{H_{TDP}} \tag{2}$$

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

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.

C. 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 [14]. Therefore, the federation complexity notion considers the following metrics.

- **Number of Training Rounds**. More rounds increase client/server computation and communication energy.
- **Dataset Size**. Larger datasets demand more memory, compute, and energy [17].
- **Model Size**. Large models typically require more computational resources and time to process each iteration, which results in higher energy consumption [17] 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 also introduce a communication overhead.
- **Number of Clients**. The more clients participate in the federation, the more energy is used [14] for i) training, ii) aggregation, and iii) communication, and thus, the more CO_2eq are emitted.
- **Client Selection Rate**. The larger this percentage, the larger the communication overhead from the uplink communication, and the larger the CO_2eq emissions.
- **Number of Local Training Rounds**. The higher the number of local training rounds, the higher the computational overhead on the client's side and the higher the energy consumption [17], [14].

D. Additional Pillars of Trustworthy FL

The six pillars defined by Sánchez et al. [5] together with the new one cover the seven requirements for trustworthy AI defined by the AI-HLEG [2] and constitute a comprehensive taxonomy.

1) Privacy: While FL offers inherent privacy benefits, it assumes honest behavior from participants. To mitigate risks from curious or malicious actors, this pillar considers four aspects: adoption of privacy-preserving methods, metrics for quantifying information leakage, and the likelihood of inferring private data from client updates.

2) Robustness: Ensuring robustness is essential to protect AI systems from adversarial threats and failures. This includes: resilience to adversarial attacks, robustness of hardware/software used by participants, and reliability of FL algorithm performance and customization.

3) Fairness: Fairness in FL is challenged by heterogeneous client data. This pillar includes client selection fairness, group-level fairness (non-discrimination), and individual-level fairness through performance alignment and label distribution balance across clients.

4) Explainability: Explainability ensures transparency in AI systems. It includes two aspects: intrinsic model interpretability and post-hoc explainability methods, which are particularly important in FL due to privacy constraints that limit access to raw data.

5) Accountability: Accountability is addressed through FactSheet Completeness, documenting the ML pipeline, and Monitoring, which ensures participants adhere to procedural and architectural standards throughout the FL model lifecycle.

6) Federation: This pillar tackles challenges in FL coordination, such as communication overhead, resource limitations, and client heterogeneity. Key aspects include client/model management and the design of optimization algorithms to ensure stable and efficient training.

IV. 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), nonfunctional 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

TABLE II: 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]

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.
- 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.
- PC_3: The metrics calculations can occur at the client's local devices, the central server, or collaboratively between both.

A. Sustainability Pillar: Notions and Metrics

TABLE II shows the notions and metrics explained in Section III 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 [15] 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 II. Besides, for the calculation of *Hardware Efficiency metrics*, the most popular benchmarking software for processors is PassMark [21]. 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 [5].

B. 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 II).
- *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 [22].
- *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 II and the metrics belonging to the remaining six pillars proposed in [5]. 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 post-training, by various participants within the federation, be it clients or servers. Once the normalized metric outputs are determined, they are assigned weights 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.

C. Algorithm Deployment

Once designed, the algorithm was implemented and deployed in a well-known FL framework called FederatedScope [23]. After the deployment, the following steps show how the sustainability and trustworthy FL scores are calculated.

1) *Setup*: FederatedScope initializes clients, server, and the trustworthiness algorithm using the given configuration, populating the FactSheet with pre-training metrics.
2) *Model Broadcast*: The server sends the global model to selected clients.
3) *Local Training*: Clients train locally and track sustainability metrics using CodeCarbon.
4) *Report Emissions*: Clients report hardware and energy grid info, stored in the emissions file.
5) *Model Sharing*: Clients send updated model parameters to the server.
6) *Aggregation*: The server securely aggregates client updates.
7) *Evaluation*: Clients evaluate the model and trigger metric computation.
8) *Iteration*: Steps 2–7 repeat for all training rounds.
9) *Finalize Results*: Final evaluation results are added to the FactSheet.
10) *Trust Score*: The algorithm computes the final trust score and generates the report.

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

V. 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.

A. 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 III. In the following experiments, 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.

Algorithm 1 Training in FederatedScope (Reduced)

```
Input: $N$ clients, sample size $m$, server $S$, iterations $T$, initial model $\overline{w}(0)$, config $C$, FederatedTrust manager $ft$
Output: Evaluation results, trust report, carbon estimates
$S$ shares client IDs and config $C$ with $ft$; $ft$ initializes FactSheet and metrics
$S$ requests class distribution; $ft$ builds emission file and class map
for each client $i \in [N]$ do
    Client $i$ reports class stats to $ft$
end for
for $t = 0$ to $T$ do
    $S$ samples $m$ clients $D(t)$ and notifies $ft$
    $S$ broadcasts model $\overline{w}(t)$
    for each client $i \in D(t)$ do
        Track emissions, train locally, update $ft$, send $w_i(t+1)$ to $S$
    end for
    $S$ tracks emissions, aggregates updates to $\overline{w}(t+1)$, updates $ft$
end for
$S$ sends final model $\overline{w}'$ to all clients
for each client $i \in [N]$ do
    Evaluate $\overline{w}'$ locally, send results to $S$
end for
$S$ aggregates results, sends to $ft$; $ft$ finalizes report and trust score
```

TABLE III: 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

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 IV, 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.

TABLE IV: 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

2) High Carbon Intensity and Low Hardware Efficiency: UC B illustrates a worst-case scenario with inefficient hardware, a highly complex federation, and carbon-intensive electricity grids, resulting in substantial CO_2eq emissions. The server uses an Intel Xeon W-2104 CPU (power performance 51.67), and all 1000 clients use AMD FX-9590 CPUs (power performance 30.76), leading to a hardware efficiency score of 0.01. The federation involves 1000 global and 90 local rounds, with a complexity score of 0.13. The server is located in South Africa (709g CO_2/kWh), with half of the clients in Kosovo (769g) and the other half in Gambia (700g), averaging 734.5g and resulting in a carbon intensity 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 sustainability with inefficient hardware and carbon-intensive electricity grids in combination with a complex federation.

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.

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, with a 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 score is 0.53, similar to UC C.

B. Effectiveness Evaluation

Nevertheless, validating the calculated sustainability pillar, which 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.

TABLE V: 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

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 V. 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 V 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. Both teams intend to conduct the training process at their local facilities.

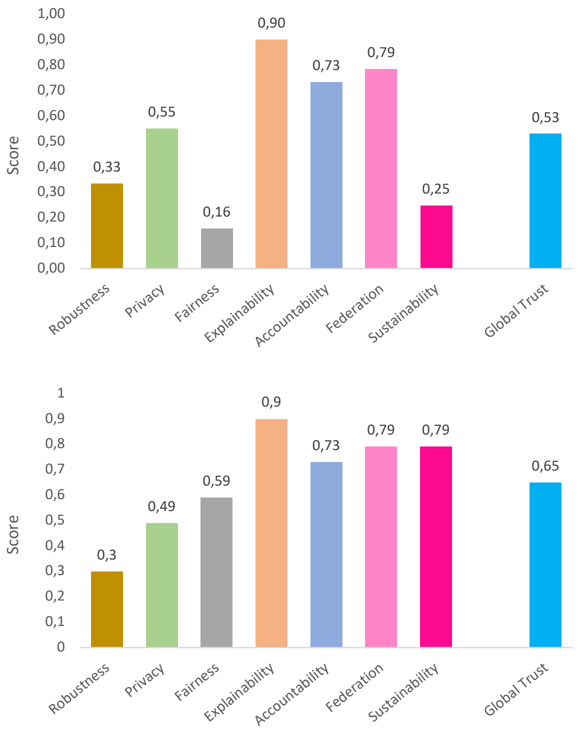

Fig. 2: Results of Evaluation of the Proposed Algorithm for Proposal A (top) and Proposal B (bottom)

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

The results of the system, as depicted in Fig. 2, 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 VI, 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.

TABLE VI: Sustainability Pillar and Notion Scores for Two Proposals

Metric	*Proposal A*	*Proposal B*
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

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 VI 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 FL systems and provides valuable support in decision-making. Moreover, it offers practical insights for the sustainable design and optimization of FL-based network management models, where energy efficiency, resource heterogeneity, and environmental constraints must be jointly considered.

VI. Discussion

The influence of individual metrics on CO_2eq emissions remains uncertain, yet all are currently weighted equally. For instance, training rounds and client count receive the same weight, despite differing environmental impacts. Hardware efficiency is currently estimated only from CPU/GPU benchmarks (e.g., PassMark), excluding RAM and embodied emissions from manufacturing, while indirect factors such as homomorphic encryption add further overhead.

Another limitation is the lack of reliable ground truth for emissions in FL. While direct measurement is difficult, recent testbeds with power consumption modules [24] can approximate ground truth and enable real-world validation of the framework. In addition, this study primarily focused on model complexity, overlooking communication, which prior work [25] shows contributes comparatively little to FL emissions. Future work will expand metrics to include communication overhead.

Experiments also show that electricity grid carbon intensity and device energy efficiency strongly affect sustainability. Thus, optimization algorithms that prioritize efficient devices in low-carbon regions while discouraging high-consumption nodes will be explored. Furthermore, the weights assigned to notions and metrics require refinement to better reflect their real-world importance.

Although computing sustainability metrics requires users to provide system-level factsheets, this typically involves metadata (e.g., hardware model, location) and poses limited privacy risks. Addressing these limitations will enhance the precision, coverage, and usability of sustainability assessments in FL systems.

VII. Summary and Future Work

This work addresses mechanisms in the area of critical decision-making, particularly FL for network and service management as a privacy-preserving collaborative AI paradigm. It extends the taxonomy of trustworthy FL by introducing a sustainability pillar to assess the environmental impact of FL systems. Ten metrics are defined across three notions: hardware efficiency, federation complexity, and carbon intensity of the energy grid, and integrated into an evaluation algorithm that incorporates hardware specifications and geographic energy profiles of clients and servers. Experiments indicate that FL systems with lower model complexity, more efficient devices, and cleaner energy sources achieve higher trustworthiness scores. The proposed approach provides actionable insights for optimizing FL deployments under heterogeneous resources, distributed environments, and energy constraints.

Future work will refine the weighting of notions and metrics to better reflect actual carbon impact, extend the coverage of communication and network infrastructure overhead, and address trade-offs between pillars (e.g., privacy mechanisms increasing sustainability cost). In addition, validation on physical testbeds with power measurement modules will serve as a proxy for ground truth emissions. Finally, enhancing the prototype's security, compatibility, and support for decentralized scenarios, as well as integrating metrics from the other six pillars, will further improve its comprehensiveness.

Acknowledgments

This work has been partially supported by *(a)* the Swiss Federal Office for Defense Procurement (armasuisse) with the CyberDFL project (CYD-C-2020003) and *(b)* the University of Zürich UZH.

References

[1] M. He, Z. Li, C. Liu, D. Shi, and Z. Tan, "Deployment of artificial intelligence in real-world practice: opportunity and challenge," *The Asia-Pacific Journal of Ophthalmology*, vol. 9, no. 4, pp. 299–307, 2020.

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

[3] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: a review," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–38, 2022.

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

[5] P. M. S. Sánchez, A. H. Celdrán, N. Xie, G. Bovet, G. M. Pérez, and B. Stiller, "Federatedtrust: A solution for trustworthy federated learning," *Future Generation Computer Systems*, vol. 152, pp. 83–98, 2024.

[6] Y. Shi, H. Yu, and C. Leung, "A survey of fairness-aware federated learning," *arXiv:2111.01872*, 2021.

[7] H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain, and J. Tang, "Trustworthy ai: A computational perspective," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 1, pp. 1–59, 2022.

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

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

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

[11] L. Zumtaugwald, "Algorithm to Compute the Sustainability and Trustworthiness of FL." https://github.com/Cyber-Tracer/SustainabilityFL. Accessed: 03.05.2025.

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

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

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

[15] CodeCarbon, "Codecarbon." https://codecarbon.io. Accessed: 22.02.2025.

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

[17] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," *arXiv:1910.09700*, 2019.

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

[19] S. Schlömer, T. Bruckner, L. Fulton, E. Hertwich, A. McKinnon, D. Perczyk, J. Roy, R. Schaeffer, R. Sims, P. Smith, *et al.*, "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*, pp. 1329–1356, Cambridge University Press, 2014.

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

[21] PassMark, "Thermal Design Power." https://www.cpubenchmark.net/power_performance.html, 2025. Accessed: 03.05.2025.

[22] M. Arnold, R. K. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, *et al.*, "Factsheets: Increasing trust in ai services through supplier's declarations of conformity," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 6–1, 2019.

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

[24] C. Feng, N. Huber, A. H. Celdran, G. Bovet, and B. Stiller, "A practical testbed for decentralized federated learning on physical edge devices," *arXiv preprint arXiv:2505.08033*, 2025.

[25] C. Feng, A. H. Celdrán, X. Cheng, G. Bovet, and B. Stiller, "Greendfl: a framework for assessing the sustainability of decentralized federated learning systems," *arXiv preprint arXiv:2502.20242*, 2025.