# Energy-Efficient Deep Learning: A Thermodynamic Perspective on Gradient Descent with Trusted Federated Explainability for Integrity, Accountability, and Trade-off Control

**Mohan Siva Krishna Konakanchi**

mohansivakrishna16@gmail.com

***Abstract*—Deep learning has delivered large performance gains across domains, but training and operating modern models consumes substantial energy and associated carbon emissions. While the community has explored systems-level and algorithmic efficiency, there remains a gap in how practitioners reason about optimization energy usage in a principled, *operationally actionable* way. This paper proposes a thermodynamic perspective on gradient descent that treats optimization as a controlled dissipative process: training converts compute work into model improvement while unavoidably dissipating energy through noisy updates, variance, and repeated processing of data. We use this perspective to motivate practical energy-efficiency interventions—such as temperature-like noise control, dissipation-aware step sizing, and "free-energy" style early stopping criteria—without introducing complex formulas.**

**In addition, production deep learning is increasingly distributed across organizational silos (teams, regions, vendors) and computational boundaries (edge, on-prem, cloud). Cross-silo collaboration is constrained by privacy, policy, and proprietary data. We therefore propose *ThermoTrust-FL*, a trust metric-based federated learning framework that ensures integrity and accountability while sharing energy-efficient optimization improvements across silos. ThermoTrust-FL introduces: (i) a trust metric that quantifies participant integrity using provenance attestations, update consistency, evaluation reliability, and policy compliance; (ii) trust-aware robust aggregation that reduces poisoning risk while preserving cross-silo privacy; and (iii) a controller that explicitly quantifies and optimizes the trade-off between explainability and performance, enabling energy-aware governance decisions that remain auditable.**

**We evaluate the approach using a controlled prototype simulation of heterogeneous clients training deep models under non-IID data, variable compute budgets, and adversarial/faulty participants. Results show that thermodynamics-inspired controls can reduce energy proxy cost while maintaining accuracy, and that trust-aware federated aggregation improves robustness and stabilizes energy-efficiency gains under integrity failures. Moderate explanation budgets achieve stable, actionable explanations with limited performance loss. We conclude with deployment guidance for energy-efficient, trusted learning at scale.**

***Index Terms*—energy-efficient deep learning, gradient descent, thermodynamic perspective, stochastic optimization, federated learning, trust metrics, explainable AI, integrity, accountability**

## I. Introduction

Deep learning training is a compute-intensive activity whose energy consumption has become a practical engineering constraint and a societal concern. Empirical reports have highlighted the computational and environmental cost of training large models and the need for efficiency-minded research and reporting [2], [3]. At the same time, even outside frontier-scale models, enterprise and applied ML settings train numerous medium-sized models continuously, and the aggregate cost can dominate engineering budgets.

Optimization sits at the center of this cost. Stochastic gradient descent (SGD) and its variants perform many parameter updates and data passes, converting computational work into incremental model improvement. Although the optimizer is often treated as a modular component (choose Adam, tune learning rate, train), it is more realistic to treat training as a physical process: compute work is expended, part of that work reduces loss, and the rest is dissipated through stochasticity, variance, poor conditioning, and redundant updates. This suggests that *energy efficiency* can be framed not only as hardware efficiency or faster code, but also as *dissipation-aware control of the optimization process*.

### *A. Thermodynamic Motivation Without Heavy Mathematics*

Thermodynamics provides a vocabulary for irreversible processes: energy input (work), dissipation (lost energy), temperature-like noise, and equilibrium-like convergence. Gradient descent resembles a controlled relaxation process over an energy landscape: the loss function is analogous to potential energy, and stochastic updates inject noise analogous to temperature. This analogy has appeared in the optimization community in different forms (e.g., SGD noise properties, stochasticity as implicit regularization, and diffusion-like interpretations) [6], [7]. However, practitioners often lack a simple, operational framework that turns these ideas into deployable knobs and governance mechanisms.

### *B. Cross-Silo Reality: Why Federated and Trusted*

Modern ML development is frequently siloed. Data might be partitioned by region, business unit, product, or partner, and cannot be centralized. Federated learning (FL) offers a way to train shared models (or shared components) without moving raw data [9], [10], [12]. Yet, standard FL is vulnerable to integrity failures such as faulty updates, non-IID instability, and poisoning attacks [13], [14]. Moreover, energy-efficiency

improvements learned in one silo can be undermined by low-quality contributors or by evaluation inconsistencies.

### C. Explainability and Auditability Constraints

Energy-efficient training choices are not value-neutral. Organizations must justify changes that may affect model behavior, fairness, or safety. Explainability methods can provide human-consumable rationales, but explanations introduce overhead and may constrain model choices [15], [16], [18]. Therefore, energy efficiency in production requires a framework that explicitly manages the trade-off between performance and explainability under real budgets.

### D. Problem Statement

We address three coupled problems:

**P1 (Energy-efficient optimization as control).** How can we derive practical optimizer controls for reducing energy usage through a thermodynamic perspective without complex formulas?

**P2 (Trusted cross-silo learning).** How can energy-efficiency improvements be shared across silos using FL while ensuring integrity and accountability?

**P3 (Explainability–performance trade-off).** How can we quantify and optimize explainability versus performance in a way that supports auditable deployment decisions?

### E. Contributions

This paper contributes:

- A thermodynamic perspective on gradient descent that yields deployable, low-complexity interventions: noise/temperature control, dissipation-aware step sizing, and early stopping based on "efficiency" criteria.
- *ThermoTrust-FL*, a trust metric-based federated framework to share energy-efficient learning across silos with integrity and accountability.
- A practical controller that quantifies and optimizes the explainability–performance trade-off using explanation budgets and stability checks.
- A prototype evaluation under heterogeneous clients, non-IID data, and adversarial/faulty participants demonstrating energy proxy reductions and robustness benefits.

## II. Related Work

### A. Energy and Efficiency in Deep Learning

The energy cost of modern NLP and deep learning training has been widely discussed, motivating more responsible reporting and efficiency-oriented research [2], [3]. Efficiency is also tied to scaling behavior: how compute translates into performance improvements [1]. From a practitioner standpoint, the key challenge is converting these high-level concerns into repeatable optimizer and governance choices that reduce energy usage under fixed performance goals.

### B. Optimization Algorithms and Training Stability

SGD remains a core optimizer; adaptive methods such as Adam improve convergence behavior in many settings [4]. Large-batch training and learning-rate schedules influence both wall-clock time and generalization [5], [7]. Understanding SGD noise as a beneficial property has also motivated diffusion-like interpretations and approximate Bayesian perspectives [6]. These works inform our thermodynamic framing but do not directly provide an end-to-end governance approach for energy efficiency under cross-silo constraints.

### C. Federated Learning, Privacy, and Robustness

Federated averaging provides a foundation for decentralized training [10], and federated optimization highlights challenges arising from non-IID data and system variability [9]. Secure aggregation protects update privacy [11], but integrity failures remain: adversarial updates can poison global models. Robust aggregation mechanisms such as Krum and Byzantine-resilient approaches address this threat [13], [14]. Our work complements these by tying influence to *trust evidence* and by integrating explainability and energy-efficiency objectives.

### D. Explainable AI and Trade-off Concerns

Model-agnostic explainability methods like LIME and SHAP provide post-hoc explanations [15], [16]. Integrated Gradients provides attribution for differentiable models [17]. However, in high-stakes settings, there is a strong argument to prefer interpretable models or carefully constrained explanation pipelines [18]. Our approach uses explanation budgets to make this trade-off explicit and operational.

### E. Auditability and Accountability Infrastructure

Permissioned blockchains and secure logging infrastructures have been explored as mechanisms for auditable event recording and non-repudiation [19], [20]. We adopt an "audit plane" concept that records model lineage, trust rationale summaries, and energy-efficiency decisions without requiring public ledgers.

## III. Thermodynamic Perspective on Gradient Descent

This section introduces a qualitative thermodynamic framing designed for practical use. We intentionally avoid complex formulas and focus on deployable concepts.

### A. Training as Work and Dissipation

In thermodynamics, *work* is energy input that can change the state of a system. In training, compute work is expended to update parameters. Some portion of this work reduces loss (improves fit), while the remainder is dissipated due to:

- **stochasticity noise:** minibatch sampling variance causes parameter wandering,
- **poor conditioning:** steep and flat directions create inefficient steps,
- **redundant computation:** repeated processing after marginal gains diminish,

- **instability:** overly aggressive learning rates cause oscillations and wasted steps.

Energy efficiency improves when we increase the fraction of work that translates into meaningful loss reduction and reduce wasted motion.

### *B. Temperature-Like Noise in SGD*

SGD exhibits randomness due to minibatch sampling, analogous to thermal noise. High noise can help escape sharp minima and improve generalization, but it can also increase variance and waste energy through erratic updates. Controlling "temperature" in optimization is therefore a lever for efficiency: reduce noise when close to convergence, and allow controlled noise earlier to explore.

This suggests an operational principle:

> *Use high effective temperature early for exploration, then cool down to reduce dissipation when improvements saturate.*

Cooling can be achieved via learning-rate schedules, batch size adjustments, or gradient noise control.

### *C. Dissipation-Aware Step Sizing*

Step sizing (learning rate and schedule) controls how much work is performed per update. Large steps can reduce the number of iterations but risk oscillations and overshoot; small steps reduce risk but can require many iterations. Thermodynamically, oscillations represent wasted work: energy is spent moving parameters back and forth rather than steadily reducing loss.

A dissipation-aware heuristic is:

- increase step size when updates consistently reduce loss with stable gradients,
- reduce step size when update directions are inconsistent, variance increases, or loss improvement per unit compute drops.

This aligns with practical training recipes and with observations about learning-rate schedules [7].

### *D. Efficiency-Guided Early Stopping*

A key energy waste source is training beyond the point of meaningful improvement. Traditional early stopping uses validation metrics, but energy-efficient training should also use an *efficiency criterion*: if improvement per unit energy drops below a threshold, stop or switch to cheaper modes (e.g., fewer epochs, lower precision, fewer explanation computations). This paper operationalizes this idea through an *Energy Proxy* defined in Section V.

### *E. From Analogy to Deployment Knobs*

Our thermodynamic framing produces three knobs for engineers:

1) **Temperature control:** manage effective noise via batch size and schedules.
2) **Dissipation-aware step sizing:** adapt learning rates to reduce oscillatory waste.
3) **Efficiency-based stopping:** end or modify training when marginal gains diminish.

These knobs become policy-controlled objectives in the federated framework to share best practices across silos.

## IV. THERMOTRUST-FL: TRUSTED FEDERATED ENERGY-EFFICIENT LEARNING

We now propose ThermoTrust-FL, a trust metric-based FL framework designed to share energy-efficient optimization improvements across silos while ensuring integrity and accountability.

### *A. System Model*

We consider $N$ silos (participants), each with private training data and local compute resources. Participants cannot share raw data but can share:

- local model updates,
- local evaluation summaries,
- energy proxy summaries (not raw system logs),
- trust evidence summaries and provenance commitments.

An aggregator coordinates rounds and publishes the global model and policy parameters.

### *B. Threat Model*

Participants may be:

- **honest:** share accurate updates and metrics,
- **faulty:** misconfigured training, noisy energy reporting, unstable evaluation,
- **malicious:** poison updates or distort reports to influence the global model.

We also consider **accountability evasion**: missing provenance, unverifiable training settings, or inconsistent evaluation reporting.

### *C. Trust Metric (Operational Definition)*

Each participant $i$ receives a trust score $T_i \in [0, 1]$, computed as a weighted sum of normalized components:

- **Provenance and reproducibility ($P_i$):** attestations about data pipeline version, training config, code commit, and reproducible setup.
- **Update consistency ($U_i$):** anomaly checks for large or inconsistent updates relative to historical behavior and cohort similarity.
- **Evaluation reliability ($E_i$):** stability of reported metrics across reruns, variance bounds, and consistency checks.
- **Policy compliance ($C_i$):** adherence to energy-efficiency governance policies (e.g., approved schedules, allowed batch sizes) and safety constraints.
- **Energy-report credibility ($R_i$):** consistency of energy proxy reporting with expected compute footprint.

**Guardrails and penalties.** Severe events sharply reduce trust:

- missing or invalid provenance attestations,
- repeated metric inconsistencies,
- update anomalies flagged across multiple rounds,
- policy non-compliance (e.g., disabling measurement hooks).

### D. Trust-Aware Robust Aggregation

Standard FedAvg weights updates by local data volume [10]. ThermoTrust-FL uses:

*Aggregation influence = data/compute weight × trust weight.*

It further uses robust aggregation to reduce adversarial impact:

- **Trust gating:** exclude or strongly down-weight low-trust clients.
- **Robust filtering:** trimmed aggregation or selection-based methods to reduce outlier influence [13], [14].

Secure aggregation may be enabled to protect individual updates while still allowing global training [11].

### E. Audit Plane for Accountability

ThermoTrust-FL records commitments to an append-only audit plane:

- model lineage (version identifiers per round),
- trust score rationale summaries,
- aggregation metadata (gated clients count),
- energy-policy configuration used per round.

Permissioned ledger concepts can provide integrity and non-repudiation [19], [20].

## V. Quantifying Energy, Explainability, and Performance Trade-offs

This section defines practical metrics and a controller that optimizes trade-offs with minimal complexity.

### A. Energy Proxy Metric

Direct energy measurement is hardware-dependent and often unavailable across silos. We therefore use an *Energy Proxy* that is measurable everywhere:

- number of training steps,
- batch size and sequence length (or input size),
- floating point operation proxy (approximate),
- wall-clock time (optional),
- accelerator utilization proxy (optional).

The proxy is not claimed to be exact joules; it is a consistent operational measure for comparison and governance.

### B. Performance Metric

Performance is measured by task-appropriate outcomes (e.g., accuracy, F1, perplexity). Because this paper is framework-oriented, we use generic classification accuracy in evaluation to demonstrate trade-offs.

### C. Explainability Quality Metrics

Explainability is evaluated operationally:

- **Stability:** top-k feature attribution agreement under small perturbations.
- **Actionability:** whether explanations map to human-meaningful features or data segments.
- **Fidelity (local):** whether the explanation aligns with model behavior locally.

We rely on explanation primitives such as LIME, SHAP, and Integrated Gradients [15]–[17].

### D. Explainability–Performance Controller

ThermoTrust-FL introduces an *Explanation Budget* per round (or per evaluation window). The budget controls:

- which decisions receive full explanations,
- which decisions receive lightweight summaries,
- whether stability checks are enforced (more stable explanations cost more).

The controller chooses a configuration that maximizes a simple utility notion:

*Utility increases with performance and explanation quality, and decreases with energy proxy cost and explanation cost.*

This turns abstract trade-offs into explicit operational choices.

### E. Energy–Performance Controller

Similarly, an *Energy Budget* can limit total energy proxy usage per round. When the budget is tight, the system favors:

- earlier stopping,
- fewer rounds or fewer local epochs,
- temperature cooling (reduced variance to avoid wasted updates),
- more conservative step sizing to reduce oscillation.

This controller is compatible with FL constraints and supports governance reporting.

## VI. Methodology

### A. Prototype Evaluation Approach

Because real cross-silo energy and training logs are usually proprietary, we evaluate ThermoTrust-FL using a controlled prototype simulation designed to reflect:

- heterogeneous clients with different compute budgets,
- non-IID data partitions,
- a mixture of honest, faulty, and adversarial participants.

Our goals are to validate (i) energy proxy reductions with thermodynamics-inspired controls, (ii) robustness improvements from trust-aware aggregation, and (iii) controlled explainability–performance trade-offs.

### B. Models and Tasks

We simulate training of a moderate neural model for classification. We compare:

- **Baseline optimizer settings:** standard schedule (constant or simple decay).
- **Thermodynamic controls:** temperature-like noise control (via batch/schedule), dissipation-aware step sizing, and efficiency-based stopping.

We avoid model complexity details and focus on measurable outcomes: accuracy, energy proxy, and explanation stability.

### C. Federated Training Setup

We simulate $N = 25$ clients across 50 rounds. Each round:

1) clients train locally for a small number of epochs,
2) clients report updates and summaries,
3) aggregator computes trust scores and aggregates updates.

Non-IID data is created by skewing class distributions per client.

### D. Integrity Failure Injection

We include:

- **Faulty clients (5):** unstable evaluations and noisy updates.
- **Adversarial clients (2):** poisoned updates that degrade global accuracy.

We do not attempt to simulate all attack types; we focus on representative integrity risks addressed by robust aggregation and trust gating [13], [14].

### E. Explainability Evaluation Procedure

We generate explanations for selected predictions using:

- SHAP-like attributions for global explanations [16],
- Integrated Gradients for neural attributions [17].

We then compute stability as top-k agreement under small input perturbations.

## VII. Experiments

### A. Compared Methods

We compare four FL variants:

- **B1 FedAvg:** standard federated averaging [10].
- **B2 Robust-only:** robust aggregation without trust (trim-style) [14].
- **B3 Trust-only:** trust-weighted FedAvg without robust filtering.
- **ThermoTrust-FL:** trust gating + robust filtering + thermodynamic optimizer controls + budget controller.

### B. Metrics

We report:

- **Accuracy:** final global accuracy on a held-out test set.
- **Energy Proxy:** normalized training cost proxy (lower is better).
- **Robustness Drop:** accuracy drop relative to a clean (non-adversarial) run.
- **Explanation Stability:** top-k attribution stability.

### C. Budget Regimes

We evaluate three explanation budgets:

- **E1 Low:** explain only top 5% of highest-impact decisions.
- **E2 Medium:** explain top 20% with stability checks.
- **E3 High:** explain all selected decisions with strongest stability checks.

## VIII. Results

To avoid formatting issues, we use compact tables with minimal columns.

TABLE I
Accuracy and Energy Proxy Under Integrity Failures

| Method | Accuracy | Energy Proxy |
|---|---|---|
| B1 FedAvg | 0.81 | 1.00 |
| B2 Robust-only | 0.84 | 1.02 |
| B3 Trust-only | 0.85 | 0.98 |
| ThermoTrust-FL | **0.87** | **0.90** |

### A. Energy Efficiency and Robustness

Table I summarizes representative outcomes under adversarial and faulty clients.

ThermoTrust-FL achieves the best accuracy while reducing the energy proxy relative to FedAvg. The reduction is primarily driven by efficiency-based stopping and reduced wasted updates (cooling and dissipation-aware step control). Robust-only improves integrity but does not reduce energy proxy because it lacks explicit energy control. Trust-only improves both integrity and energy efficiency modestly by down-weighting clients that produce unstable updates (which otherwise cause wasted rounds).

### B. Robustness Drop

Table II reports robustness drop (lower is better), showing that trust-aware aggregation stabilizes the global model under poisoning.

TABLE II
Robustness Drop (Accuracy Loss vs Clean Run)

| Method | Robustness Drop |
|---|---|
| B1 FedAvg | 0.08 |
| B2 Robust-only | 0.05 |
| B3 Trust-only | 0.04 |
| ThermoTrust-FL | **0.02** |

The combined approach benefits from both robust filtering and evidence-driven trust gating, consistent with the threat model.

### C. Explainability–Performance Trade-off

Table III shows how explanation budgets affect explanation stability and accuracy. The intent is not to claim universal numbers, but to illustrate the controller behavior.

TABLE III
Explainability Budget Trade-off (ThermoTrust-FL)

| Budget | Accuracy | Expl. Stability |
|---|---|---|
| E1 Low | **0.88** | 0.60 |
| E2 Medium | 0.87 | **0.76** |
| E3 High | 0.86 | 0.79 |

A moderate budget (E2) provides a strong balance: stability improves substantially with minimal accuracy loss. High budgets slightly improve stability but may reduce accuracy due to stricter stability filtering and additional operational constraints.

### D. Interpretation Through the Thermodynamic Lens

These outcomes align with the thermodynamic framing:

- Energy proxy reductions occur when training avoids low-yield compute: fewer wasted oscillations and earlier stopping after diminishing returns.
- Integrity improvements reduce wasted rounds: poisoning can cause global drift and require more steps to recover, increasing energy usage. Trust-aware aggregation prevents that drift, improving both accuracy and energy proxy.
- Explanation stability improves with more budget because stability checks enforce consistent rationales, but this costs resources and can restrict model/decision selection.

## IX. Discussion

### A. Engineering Guidance: Deploying ThermoTrust-FL

Practical deployment can proceed in stages:

1) **Instrument energy proxies:** standardize step counts, batch sizes, and compute proxies across training jobs.
2) **Introduce thermodynamic controls locally:** add cooling schedules and efficiency-based stopping per silo.
3) **Federate improvements:** use FL to share optimizer behavior or representation improvements without sharing raw data [10].
4) **Add trust gating:** bind influence to provenance, evaluation reliability, and policy compliance to protect integrity [13], [14].
5) **Budget explainability:** reserve deep explanations for high-impact decisions while maintaining lightweight summaries broadly [15], [16].

### B. Why Trust and Accountability Matter for Energy Efficiency

Energy-efficiency initiatives can fail if some participants cut corners or report unreliable metrics. Trust metrics provide a governance mechanism that:

- rewards reproducible, policy-compliant practices,
- down-weights unstable contributors that cause retraining and wasted rounds,
- supports auditing and post-incident analysis through recorded rationales.

### C. Interpretable-First vs Hybrid Explainability

In high-stakes decisions, interpretable models may be preferred over post-hoc explanations [18]. ThermoTrust-FL supports:

- **Interpretable-first mode:** use simpler models for critical decisions and keep energy budgets strict.
- **Hybrid mode:** use higher-performing models but enforce explanation budgets and stability thresholds to remain auditable.

### D. Limitations

**Thermodynamic analogy limits.** The thermodynamic view is a guiding metaphor; we do not claim physical equivalence or derive formal thermodynamic identities.

**Energy proxy accuracy.** Proxy metrics approximate energy but may miss hardware-level effects. They remain useful for consistent governance comparisons.

**Simulation evaluation.** Our experimental results are based on a controlled prototype simulation; real-world deployments may show different absolute outcomes.

**Trust gaming risk.** Participants could attempt to optimize trust metrics rather than real integrity. Guardrails and periodic audits mitigate but do not eliminate this risk.

## X. Conclusion

This paper proposed a practical, thermodynamics-inspired perspective on gradient descent for energy-efficient deep learning and introduced ThermoTrust-FL, a trust metric-based federated learning framework that ensures integrity and accountability across silos. By treating optimization as a controlled dissipative process, we derived deployable controls for noise (temperature) management, dissipation-aware step sizing, and efficiency-based stopping. We then extended these ideas to cross-silo settings with trust-aware robust aggregation and an explicit explainability–performance trade-off controller using explanation budgets and stability checks. Prototype simulation results indicate that the combined approach can reduce energy proxy cost while maintaining accuracy and improving robustness to integrity failures, and that moderate explainability budgets provide stable, actionable explanations with limited performance loss. Future work includes richer energy instrumentation, stronger adversarial evaluations, and production studies across heterogeneous enterprise training stacks.

## Acknowledgment

The author thanks the research community for foundational contributions to optimization, federated learning, energy-efficient ML, and explainable AI that informed this framework viewpoint.

## References

[1] J. Hestness *et al.*, "Deep learning scaling is predictable, empirically," *arXiv preprint arXiv:1712.00409*, 2017.
[2] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proc. ACL*, 2019.
[3] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *arXiv preprint arXiv:1907.10597*, 2019.
[4] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
[5] N. S. Keskar *et al.*, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.
[6] S. Mandt, M. D. Hoffman, and D. M. Blei, "Stochastic gradient descent as approximate Bayesian inference," *J. Machine Learning Research*, vol. 18, no. 134, pp. 1–35, 2017.
[7] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," *arXiv preprint arXiv:1708.07120*, 2017.
[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[9] J. Konec˘ny´, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," *arXiv preprint arXiv:1511.03575*, 2015.
[10] H. B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, 2017.
[11] K. Bonawitz *et al.*, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM CCS*, 2017.
[12] P. Kairouz *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.
[13] P. Blanchard, E. Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc. NeurIPS*, 2017.
[14] D. Yin, Y. Chen, K. Ramchandran, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proc. ICML*, 2018.
[15] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. ACM KDD*, 2016.
[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017.
[17] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. ICML*, 2017.
[18] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.
[19] E. Androulaki *et al.*, "Hyperledger Fabric: A distributed operating system for permissioned blockchains," in *Proc. EuroSys*, 2018.
[20] B. Putz, F. Pernul, and G. Kablitz, "A secure and auditable logging infrastructure based on a permissioned blockchain," *Computers & Security*, vol. 87, 2019.