Gaming and Cooperation in Federated Learning: What Can Happen and How to Monitor It

Anonymous authors
Paper under double-blind review

Abstract

The success of federated learning (FL) ultimately depends on how strategic participants behave under partial observability, yet most formulations still treat FL as a static optimization problem. We instead view FL deployments as governed strategic systems and develop an analytical framework that separates welfare-improving behavior from metric gaming. Within this framework, we introduce indices that quantify manipulability, the price of gaming, and the price of cooperation, and we use them to study how rules, information disclosure, evaluation metrics, and aggregator-switching policies reshape incentives and cooperation patterns. We derive threshold conditions for deterring harmful gaming while preserving benign cooperation, and for triggering auto-switch rules when early-warning indicators become critical. Building on these results, we construct a design toolkit including a governance checklist and a simple audit-budget allocation algorithm with a provable performance guarantee. Simulations across diverse stylized environments and a federated learning case study consistently match the qualitative and quantitative patterns predicted by our framework. Taken together, our results provide design principles and operational guidelines for reducing metric gaming while sustaining stable, high-welfare cooperation in FL platforms.

1 Introduction

1.1 Motivation

Federated learning (FL) enables multiple organizations to train a shared model without moving data, and adoption has been growing through consortia and platform-based collaborations. As cross-organizational data collaboration expands, keeping data local while jointly training models has become a leading alternative to traditional data pooling, especially when privacy, confidentiality, and regulatory constraints rule out centralization. In this setting, FL naturally supports emerging data and AI services in which participants retain control over their local data while contributing to shared intelligence that is monetized or governed through metrics, contracts, and service-level guarantees.

At the same time, the surrounding governance has not matured at the same pace. Many organizations are still in early stages of formalizing procedures for AI-enabled services, clarifying accountability, and embedding risk mitigation into day-to-day operations. In FL, these gaps are particularly consequential: joint outcomes depend on the actions of multiple parties whose internal processes are opaque to one another, yet whose rewards and reputation are coupled through shared metrics. When rewards, rankings, or access rights depend on these metrics, participants face incentives not only to improve genuine performance but also to target the metrics themselves, potentially drifting toward high-metric, low-welfare regimes.

This tension is amplified by the combination of privacy and limited observability. Privacy-enhancing technologies such as FL, differential privacy, and secure computation restrict what can be inspected or audited, and information design around metrics and scores further shapes what participants see and react to. Strong protections are desirable, but they also reduce the visibility of individual behavior and may unintentionally make harmful manipulation harder to detect. Designing FL systems therefore requires more than choosing an optimizer or aggregation rule: it requires a coordinated view of evaluation, information disclosure, rewards, audits, and participation, and of how these choices jointly shape incentives, cooperation, and stability.

This paper aims to provide such a view. We treat FL as a governed strategic system and develop a three-layer framework that (i) quantifies how design choices create room for metric gaming and cooperation, (ii) links these quantities to participation dynamics and tipping points, and (iii) maps them to practical levers such as mixed public–private evaluation, audit allocation, sanction calibration, and auto-switch rules. Our goal is not to propose yet another FL algorithm, but to offer a compact language and toolkit that helps designers reason about metric gaming, cooperation, and governance in federated environments.

1.2 Contributions

We view federated learning not merely as a distributed optimization problem but as a *governed strategic system* in which evaluation rules, information disclosure, reward and sanction mechanisms, and audit capacity jointly shape participants' incentives. Our contributions are to provide a formal language for this system, to connect that language to simple indices and thresholds that can be estimated from data, and to distill resulting design principles into an actionable toolkit.

- **Strategic formalization of federated learning.** We formalize a generic *Eval–Info–Reward–Audit* architecture for federated learning platforms (Section 3), specifying welfare, public and private metrics, participation choices, and cooperative actions (e.g., coalition formation, data sharing) within a single game-theoretic environment. This provides a common backbone on which existing robust aggregation, incentive, and privacy mechanisms can be placed.

- **Indices for manipulability, gaming, and cooperation.** On top of this backbone, we introduce three indices that summarize how a design policy π trades off metric performance and welfare: a manipulability index $M(\pi)$ that captures the best achievable metric gain per unit welfare loss under unilateral deviations, a Price of Gaming $\mathrm{PoG}(\pi)$ that measures welfare loss when a fraction of clients adopt gaming behaviors, and a Price of Cooperation PoC that quantifies the net welfare effect of coalitions (Section 4). We establish basic properties of these indices, show how they interact with simple penalty schemes, and use them to define design-time thresholds $(\alpha_{\min}, \alpha_{\text{benign}})$ that separate under-enforcement, harmful gaming regimes, and over-enforcement that discourages benign cooperation.

- **Participation dynamics, resilience, and tipping thresholds.** We couple the metric layer to a stylized participation dynamics model, defining an aggregate participation map $F(x; \pi)$ and a resilience indicator $R(\pi)$ that summarize how participation responds to policy changes (Section 5). Under this model, we characterize tipping points and domino-style exits, and show how the indices above bound the regions in which small changes in sanction strength or public-metric weight can trigger large shifts in participation. This yields interpretable thresholds and heuristic rules for maintaining participation stability while suppressing high-PoG equilibria.

- **Design toolkit for evaluation, audits, and governance.** Building on these indices and dynamics, we propose a set of design patterns for federated platforms (Section 6), including mixed public/private evaluation schemes, audit-budget allocation rules with $(1 - 1/e)$ approximation guarantees for detecting high-manipulability clients, and a governance checklist that links observable diagnostics (e.g., gaming incident rates, participation responses, coalition structures) to concrete levers (metric choice, disclosure policy, sanction strength, and audit allocation).

- **Empirical illustrations in stylized and federated settings.** Finally, we instantiate the framework in a stylized simulator and a federated Fashion-MNIST experiment (Section 7), showing how high-metric/low-welfare equilibria, participation tipping, and benign versus harmful cooperation patterns arise under different design policies. We demonstrate that our indices can be approximated using simple scenario-based experiments and log-based diagnostics, and that the proposed toolkit can detect and mitigate high-PoG regimes without collapsing participation.

2 Related Work

2.1 FL Attacks, Defenses, and Robust Aggregation

Early work on robustness in federated and distributed learning asked whether a small number of malicious or corrupted updates can derail training, leading to robust aggregation rules such as Krum, coordinate-wise median and trimmed mean, Bulyan, and Robust Federated Aggregation (RFA) (Blanchard et al., 2017; Yin et al., 2018; Guerraoui et al., 2018; Pillutla et al., 2022), along with convergence analyses under Byzantine noise and coding-based defenses (Alistarh et al., 2018; Bernstein et al., 2018; Chen et al., 2018). Subsequent studies implemented these mechanisms under realistic constraints and exposed their limits via model-poisoning, backdoor, and tail-group attacks, and proposed collusion- and Sybil-aware defenses such as FoolsGold and FLTrust (Damaskinos et al., 2019; Xie et al., 2020; Baruch et al., 2019; Fang et al., 2020; Bagdasaryan et al., 2020; Wang et al., 2020; Fung et al., 2018; Cao et al., 2020). Our work does not add another aggregation or poisoning defense; instead, we ask how the choice of rules, metrics, and audits shapes incentives for manipulation and cooperation, and how these incentives affect aggregate performance and participation even when the underlying optimization dynamics are well behaved.

2.2 Incentive Design and Differential Client Contribution

Differential contribution and reward design often start from data valuation: Shapley-value-based methods, influence functions, and learned value estimators seek fair allocations by tracing marginal contributions of data or clients (Ghorbani & Zou, 2019; Jia et al., 2019; Koh & Liang, 2017; Yoon et al., 2020; Liu et al., 2022b; Chen et al., 2023; 2024; Tastan et al., 2024). Direct incentive schemes use reputation, contracts, auctions, and blockchain-based mechanisms to induce participation and effort, while parallel work folds contribution weighting into learning and aggregation to target fairness, robustness, or resource-aware client selection (Kang et al., 2019; Zhang et al., 2021a; Tang et al., 2024; Zhang et al., 2021b; Liu et al., 2020; Mohri et al., 2019; Li et al., 2019; Lai et al., 2021; Nishio & Yonetani, 2019; Lin et al., 2023; Kim et al., 2024; Ouyang & Kuang, 2025). These approaches typically treat incentives as reward-allocation functions on fixed metrics or as modified aggregation rules; in contrast, we focus on the strategic environment induced by evaluation and audit design itself and introduce indices that quantify when such designs incentivize gaming versus genuine improvement.

2.3 Game Theory of Federated Learning and Participation Dynamics

Game-theoretic treatments of FL have emphasized coalition formation, stability, and free-riding, using hedonic games, coalition models, and cooperative-game valuations to study gaps between individually stable and globally optimal outcomes (Donahue & Kleinberg, 2021a;b; Hasan, 2021; Nagalapatti & Narayanam, 2021). Repeated-game and leader–follower analyses show how punishment strategies, contract-theoretic mechanisms, and Stackelberg formulations can deter free-riding and align server–client incentives under private information, including collusion- and Sybil-aware variants (Zhang et al., 2022; Sagduyu, 2022; Luo et al., 2023; Sarikaya & Ercetin, 2019; Hu & Gong, 2020; Le et al., 2021; Ding et al., 2020; Liu et al., 2022a; Huang et al., 2024; Byrd et al., 2022; Xiong et al., 2024). Our work shares the strategic perspective but shifts focus from specific mechanism equilibria to a metric-based language and threshold results for participation stability, tipping points, and domino exits under generic rule sets.

2.4 Metric Gaming and the Goodhart Phenomenon

Metric gaming is often summarized by Goodhart's observation that once a measure becomes a target, it ceases to be a good measure. Prior work has categorized mechanisms behind this phenomenon and documented failure modes such as reward hacking, non-scalable oversight, and goal misgeneralization (Manheim & Garrabrant, 2018; Amodei et al., 2016; Skalse et al., 2022; Everitt et al., 2021; Di Langosco et al., 2022). When metrics drive decisions and thereby change the data, the problem becomes strategic: models of strategic classification, Stackelberg interactions, and performative prediction analyze how agents respond to public classifiers and how repeated retraining interacts with distributional shifts (Hardt et al., 2016a; Brückner &

Scheffer, 2011; Perdomo et al., 2020; Mendler-Dünner et al., 2020; Miller et al., 2020). We build on these insights but specialize them to FL, focusing on how evaluation metrics, disclosure policies, and audits induce client-level incentives for metric targeting and introducing explicit indices and thresholds that distinguish welfare-improving behavior from gaming.

2.5 Design of Evaluation Information, Audits, and Sanctions

Information design around metrics and scores has been proposed as a primary tool against overfitting and gaming, via staircase leaderboards, reusable holdouts, and adaptive-data-analysis bounds for safe repeated evaluation (Blum & Hardt, 2015; Dwork et al., 2015b;c;a; Bassily et al., 2016). From the audit side, work on fairness, documentation, behavior-based testing, and distribution-shift benchmarks has developed procedures and artifacts that surface vulnerabilities beyond single scalar metrics (Kearns et al., 2018; Hardt et al., 2016b; Kim et al., 2019; Agarwal et al., 2018; Kusner et al., 2017; Mitchell et al., 2019; Gebru et al., 2021; Ribeiro et al., 2020; Kiela et al., 2021; Koh et al., 2021; Croce et al., 2020). Privacy and memorization audits—including membership inference, memorization measures, data extraction, and backdoor detection—further highlight the role of audits and sanctions in deployed systems (Shokri et al., 2017; Carlini et al., 2019; 2021; Tran et al., 2018; Wang et al., 2019; Rabanser et al., 2019; Dressel & Farid, 2018). These strands provide building blocks for evaluation and auditing but are usually studied separately from incentives and participation; our contribution is to integrate evaluation information design, audit allocation, and sanction triggers into a single FL framework with indices and thresholds linked to gaming incentives and cooperation stability.

2.6 Trade-offs Between Privacy and Incentives

Privacy in FL is commonly enforced through client-level differential privacy and secure aggregation, often combined with distillation or selective-sharing protocols to limit exposure of sensitive data (Abadi et al., 2016; McMahan et al., 2018; Geyer et al., 2017; Bonawitz et al., 2017; Erlingsson et al., 2019; Papernot et al., 2017; 2018; Shokri & Shmatikov, 2015; Gilad-Bachrach et al., 2016). Refined attack models, including gradient inversion, attribute inference, backdoor and membership attacks, have clarified the limits of these protections and the tension between strong privacy, visibility of malicious behavior, and group-level performance and fairness (Melis et al., 2019; Zhu et al., 2019; Bagdasaryan et al., 2020; Wang et al., 2020; Bagdasaryan et al., 2019; Jagielski et al., 2019; Tramer & Boneh, 2020). Recent work on proofs of correct training and bidirectional verification protocols aims to restore some auditability under privacy constraints (Jia et al., 2021; Zhang & Yu, 2022), but typically remains orthogonal to incentive design. Our framework explicitly foregrounds the trade-offs between privacy, audit signals, and incentives, and proposes design patterns—such as randomized evaluation, limited disclosure, and connectivity-based alarms—that help maintain incentive alignment and sanctionability under realistic privacy and legal constraints. While prior work typically focuses on individual mechanisms, attacks, or defenses, our contribution is to provide a platform-level framework that jointly links metric gaming, welfare loss, participation dynamics, and governance levers through explicit indices and design thresholds.

3 Federated Learning as a Strategic System: Setup and Notation

In this section we formalize federated learning (FL) as a strategic system. The goal is not a fully realistic economic model but a minimal structure that supports the indices and dynamics developed later.

3.1 Federated environment and timing

We consider a cross-silo FL environment with a finite set of clients

$$\mathcal{I} = \{1, \ldots, n\}$$

and a coordinating server (or platform). Time is discrete,

$$t = 1, 2, \ldots, T,$$

where T may be finite or infinite. At each round t, the server maintains a global model

$$\theta_t \in \mathbb{R}^d,$$

and each client i holds a local dataset D_i drawn from an unknown distribution $\mathcal{P}_i$. We write $\mathcal{P} = \{\mathcal{P}_i\}_{i\in\mathcal{I}}$ for the collection of client distributions and $P^\star$ for the target population distribution of interest (e.g., a mixture of the $\mathcal{P}_i$ or an external deployment distribution).

A generic round proceeds as follows.

1. **Broadcast:** The server broadcasts θ_t (or a transformed version) to eligible clients.
2. **Participation choice:** Each client i chooses

   $$p_{i,t} \in \{0, 1\},$$

   and, if $p_{i,t} = 1$, selects a local training and reporting behavior from a feasible action set (defined below).
3. **Local computation:** Participating client i runs a procedure $\mathsf{Train}_{i,t}$ on (D_i, θ_t) to obtain an internal update

   $$u_{i,t}^{\text{int}} \in \mathbb{R}^d$$

   (e.g., a gradient, parameter delta, or full model).
4. **Reporting:** Client i transforms $u_{i,t}^{\text{int}}$ into a reported update

   $$u_{i,t} \in \mathbb{R}^d,$$

   which may coincide with $u_{i,t}^{\text{int}}$ (honest), be perturbed (e.g., for privacy or obfuscation), or be arbitrarily chosen (e.g., malicious or collusive).
5. **Aggregation and evaluation:** The server aggregates updates from

   $$\mathcal{I}_t := \{i \in \mathcal{I} : p_{i,t} = 1\}$$

   via an aggregation rule Agg_t,

   $$\theta_{t+1} = \mathsf{Agg}_t\big(\theta_t, \{u_{i,t}\}_{i\in\mathcal{I}_t}\big),$$

   and runs an evaluation pipeline Eval_t to produce scores and signals (Section 3.3).
6. **Rewards, audits, and sanctions:** Based on evaluation outcomes and history, the server computes rewards or payments $R_{i,t}$, selects clients to audit, and applies sanctions or adjustments to future eligibility.

The tuple

$$\mathcal{E} = \big(\mathcal{I}, \mathcal{P}, P^\star, \{D_i\}_{i\in\mathcal{I}}, \{\theta_t\}_{t\geq 1}, \{\mathsf{Agg}_t\}_{t\geq 1}\big)$$

specifies the FL environment; strategic and governance structure is determined by how actions, metrics, rewards, and audits are defined on top of this environment.

3.2 Actions, outcomes, metrics, and welfare

At each round t, client i chooses a (possibly history-dependent) action

$$a_{i,t} \in \mathcal{A}_i,$$

where $\mathcal{A}_i$ is a feasible action set. We model $a_{i,t}$ as encoding three components:

- *Participation:* whether to participate, $p_{i,t} \in \{0, 1\}$;

- *Local effort and training:* a local training policy (e.g., epochs, learning rate, regularization, data selection) that determines $u_{i,t}^{\text{int}}$;
- *Reporting and manipulation:* a reporting policy that maps $(\theta_t, u_{i,t}^{\text{int}}, D_i, \text{history})$ to $u_{i,t}$ and any auxiliary reports (e.g., claimed statistics).

We write $a_t = (a_{i,t})_{i\in\mathcal{I}}$ for the joint action profile and h_t for the public history up to and including round t (e.g., past scores, sanctions, aggregate statistics). A *strategy* for client i is a mapping

$$\sigma_i : h_{t-1} \mapsto a_{i,t},$$

and $\sigma = (\sigma_i)_{i\in\mathcal{I}}$ is a strategy profile.

True outcomes and welfare. Given a sequence of actions $\{a_t\}_{t\geq 1}$ and environment $\mathcal{E}$, the induced models $\{\theta_t\}_{t\geq 1}$ yield time-dependent *true performance* (or welfare) on $P^\star$,

$$W_t(\sigma) := W(\theta_t; P^\star),$$

where W is a task-dependent welfare functional (e.g., negative risk, utility, or a composite measure). We sometimes summarize long-run welfare by

$$W(\sigma) := \Phi\big(\{W_t(\sigma)\}_{t\geq 1}\big),$$

such as steady-state or discounted average welfare.

Metrics and proxy performance. In practice, rewards, sanctions, and policy changes depend on *metrics* computed from finite evaluation sets and partial information. Let

$$M_t(\sigma) \in \mathbb{R}^k$$

denote a vector of metrics at round t (e.g., validation accuracy, per-group performance, fairness or robustness indicators), and

$$M(\sigma) := \Psi\big(\{M_t(\sigma)\}_{t\geq 1}\big)$$

an aggregate functional (e.g., time-averaged or worst-case group performance) used for decision-making.

In general $M(\sigma)$ need not coincide with $W(\sigma)$, and strategic behavior is driven by how M enters each client's payoff, not by W directly. The indices in Section 4 quantify the extent to which agents can change M without commensurate improvements in W.

Client payoffs. Each client i receives a cumulative payoff

$$U_i(\sigma) = \mathbb{E}\Big[\sum_{t=1}^{T} \delta^{t-1}\Big(R_{i,t}(\sigma) - C_{i,t}(\sigma)\Big)\Big],$$

where $R_{i,t}$ is the (possibly stochastic) reward at round t, $C_{i,t}$ denotes costs (e.g., computation, communication, privacy loss, regulatory risk, or sanctions), $\delta \in (0,1]$ is a discount factor, and the expectation is over randomness in training, evaluation, and audits. We treat rewards and costs as induced by design choices in evaluation, audits, and sanctions and focus on the qualitative structure of U_i rather than precise functional forms.

Definition 3.1 (Federated strategic system)**.** A *federated strategic system* is a tuple

$$\mathcal{G} = \big(\mathcal{E}, \{\mathcal{A}_i\}_{i\in\mathcal{I}}, \mathsf{Eval}, \mathsf{Info}, \mathsf{Reward}, \mathsf{Audit}\big),$$

where $\mathcal{E}$ is the FL environment, $\mathcal{A}_i$ is the action set for client i, Eval is the evaluation mechanism that produces metrics M_t, Info is the information-disclosure mechanism, Reward maps observable signals to rewards $R_{i,t}$, and Audit specifies audit and sanction rules. Together, these components induce a game in which clients choose strategies σ_i to maximize $U_i(\sigma)$, while the designer is interested in the induced welfare $W(\sigma)$ and metric behavior $M(\sigma)$.

This formulation makes explicit that algorithmic components (e.g., Agg_t and local training) are only one part of the system; evaluation, information, reward, and audit mechanisms are equally central in determining strategic behavior.

3.3 Observation, reward, and audit channels

Design choices in our framework are expressed through three coupled channels: observation (what is measured), information (what is disclosed), and enforcement (how rewards and audits respond).

Evaluation and observation. At each round t, the evaluation mechanism Eval_t maps the current and candidate models, together with internal evaluation data, to raw scores, for example:

- a global score S_t^{glob} measuring overall performance on a held-out set;
- per-client or per-group scores $S_{i,t}^{\text{loc}}$ (e.g., performance on client-i's data or on a cluster of similar clients);
- challenge outcomes $C_{i,t}$ from randomized or private tests not fully observable to clients.

We collect these into a raw observation vector

$$O_t = \big(S_t^{\text{glob}}, \{S_{i,t}^{\text{loc}}\}_{i\in\mathcal{I}}, \{C_{i,t}\}_{i\in\mathcal{I}}\big),$$

which may depend on the entire history of actions and models up to round t.

Information disclosure. The information mechanism Info_t selects which components of O_t to disclose, and at what granularity, to each client and to the public. We denote by

$$Z_{i,t} = \mathsf{Info}_t(i, O_t, h_{t-1})$$

the signal revealed to client i at round t. This may include public leaderboard entries (possibly perturbed or delayed), private feedback about $S_{i,t}^{\text{loc}}$ or $C_{i,t}$, and aggregate statistics about other clients or coalitions. Clients choose actions based on anticipated and observed signals $Z_{i,t}$, not on the full observation O_t.

Rewards and sanctions. The reward mechanism maps disclosed signals and history into transfers and penalties. For each client i and round t,

$$R_{i,t} = \mathsf{Reward}_t(i, Z_{i,t}, h_{t-1}),$$

and the audit–sanction mechanism

$$\mathsf{Audit}_t : (O_t, h_{t-1}) \mapsto \Big(A_t, \{P_{i,t}\}_{i\in\mathcal{I}}\Big)$$

selects a subset $A_t \subseteq \mathcal{I}$ to audit under a budget constraint (e.g., $\mathbb{E}[|A_t|] \le B_t$) and specifies sanctions $P_{i,t}$ (e.g., fines, exclusion, score resets). Sanctions may also be triggered automatically by threshold violations in O_t or by connectivity-based alarms constructed from recent histories.

Design policies. For many of our results, it is convenient to view the tuple

$$\pi = (\mathsf{Eval}, \mathsf{Info}, \mathsf{Reward}, \mathsf{Audit})$$

as a *design policy* chosen by the system designer. Different policies π induce different games $\mathcal{G}(\pi)$ and thus different equilibria and welfare outcomes.

In the next section, we use this setup to define indices such as the Manipulability Index and the Prices of Gaming and Cooperation, which quantify how a given design policy π shapes the gap between metric performance $M(\sigma)$ and true welfare $W(\sigma)$ and how strongly it incentivizes gaming versus cooperation.

4 Metric Layer: Indices for Gaming and Cooperation

Building on the strategic formulation in Section 3, we now introduce indices that quantify (i) how much room a design policy leaves for metric gaming and (ii) how costly gaming and cooperation are for collective welfare. These indices form the *metric layer* of our framework: they ignore the precise structure of training and aggregation and focus on how evaluation, information, and incentives shape the relationship between metric performance and true welfare.

Throughout this section, we fix an environment $\mathcal{E}$ and a design policy $\pi = (\mathsf{Eval}, \mathsf{Info}, \mathsf{Reward}, \mathsf{Audit})$, and write $\mathcal{G}(\pi)$ for the induced federated strategic system (Section 3.2).

4.1 Decomposing welfare and metric responses

We first formalize how unilateral deviations by a single client affect welfare and metrics.

Definition 4.1 (Local welfare and metric responses)**.** Let σ be a strategy profile in $\mathcal{G}(\pi)$, and let σ_i' be an alternative strategy for client i. We define

$$\Delta W_i(\sigma_i' \mid \sigma) := W(\sigma_i', \sigma_{-i}) - W(\sigma), \qquad \Delta M_i(\sigma_i' \mid \sigma) := M(\sigma_i', \sigma_{-i}) - M(\sigma),$$

as the change in true welfare and in the aggregate metric induced by client i deviating from σ_i to σ_i' while all other clients keep σ_{-i}.

We interpret $\Delta W_i > 0$ as welfare-improving and $\Delta W_i < 0$ as welfare-harming, with an analogous interpretation for ΔM_i. Our primary interest is in deviations that *improve* the metric while leaving welfare unchanged or worse.

Definition 4.2 (Metric-gaming deviation)**.** A deviation $\sigma_i' \neq \sigma_i$ at profile σ is a *metric-gaming deviation* if

$$\Delta M_i(\sigma_i' \mid \sigma) > 0 \quad \text{and} \quad \Delta W_i(\sigma_i' \mid \sigma) \leq 0.$$

It is *strongly gaming* if $\Delta W_i(\sigma_i' \mid \sigma) < 0$.

For equilibria, we often focus on steady-state profiles $\sigma^\dagger$ (e.g., stationary or long-run equilibria) of $\mathcal{G}(\pi)$, but the indices below apply equally to transient profiles.

4.2 Manipulability index

Intuitively, the manipulability of a design policy π measures how much a client can improve the metric without improving welfare, relative to the scale of attainable welfare improvements.

Definition 4.3 (Local manipulability at a profile)**.** Let σ be a strategy profile in $\mathcal{G}(\pi)$. The *local manipulability* at σ is

$$\mathcal{M}(\sigma; \pi) := \sup_{i \in \mathcal{I}} \sup_{\sigma_i' \in \mathcal{A}_i} \frac{\left[\Delta M_i(\sigma_i' \mid \sigma)\right]_+}{\left[\Delta W_i(\sigma_i' \mid \sigma)\right]_+ + \varepsilon},$$

where $[x]_+ = \max\{x, 0\}$ and $\varepsilon > 0$ is a small normalization constant that prevents division by zero when no welfare-improving deviation exists.

The numerator is the largest metric gain a client can obtain by deviating from σ, and the denominator normalizes by the scale of possible welfare improvements. Heuristically, small $\mathcal{M}(\sigma; \pi)$ means metric gains are only available when accompanied by comparable welfare gains, while large $\mathcal{M}(\sigma; \pi)$ means the metric can move substantially in directions that barely move welfare. The additive ε matters near welfare optima, where no strictly welfare-improving deviations exist.

Definition 4.4 (Manipulability index of a design policy)**.** For a design policy π and a reference class of profiles Σ^{ref} (e.g., steady-state equilibria), the *manipulability index* is

$$\mathcal{M}(\pi) := \sup_{\sigma \in \Sigma^{\mathrm{ref}}} \mathcal{M}(\sigma; \pi).$$

When Σ^{ref} consists of equilibria, $\mathcal{M}(\pi)$ measures how much the metric can be locally moved without commensurate welfare improvement around points actually induced by the policy.

Proposition 4.5 (Zero manipulability and local alignment)**.** *Suppose that for a design policy π and reference class Σ^{ref} we have $\mathcal{M}(\pi) = 0$. Then for every $\sigma \in \Sigma^{\mathrm{ref}}$, every client i, and every deviation $\sigma_i' \in \mathcal{A}_i$,*

$$\Delta M_i(\sigma_i' \mid \sigma) > 0 \quad \Rightarrow \quad \Delta W_i(\sigma_i' \mid \sigma) > 0.$$

In particular, there are no metric-gaming deviations at any profile in Σ^{ref}.

Remark 4.6. Large values of $\mathcal{M}(\pi)$ do not by themselves guarantee harmful gaming, but they quantify the capacity of the metric to move independently of welfare. In later sections we combine $\mathcal{M}(\pi)$ with audit and sanction rules to reason about when this capacity is actually exploited.

4.3 Price of Gaming

Manipulability describes local capacity for gaming; we next quantify the realized welfare loss when gaming occurs under a given policy. Inspired by the Price of Anarchy, we define the *Price of Gaming* as the welfare gap between an idealized aligned behavior and a gaming equilibrium, normalized by the aligned welfare.

We distinguish two benchmark profiles:

- A *welfare-aligned* benchmark σ^{align}, representing the best outcome achievable under π when clients are constrained to actions $\mathcal{A}_i^{\mathrm{align}} \subseteq \mathcal{A}_i$ that exclude metric-gaming behaviors (e.g., truthful reporting and genuine effort).
- A *gaming equilibrium* σ^{game}, representing an equilibrium of $\mathcal{G}(\pi)$ when clients may use the full action sets $\mathcal{A}_i$, including manipulative behavior.

Definition 4.7 (Price of Gaming)**.** Let σ^{align} and σ^{game} be as above, and suppose $W(\sigma^{\mathrm{align}}) > 0$. The *Price of Gaming* under design policy π is

$$\mathrm{PoG}(\pi) := \frac{W(\sigma^{\mathrm{align}}) - W(\sigma^{\mathrm{game}})}{W(\sigma^{\mathrm{align}})}.$$

By construction, $\mathrm{PoG}(\pi) \in [0, 1]$ when $W(\sigma^{\mathrm{game}}) \geq 0$. A value of $\mathrm{PoG}(\pi) = 0$ indicates that gaming does not reduce welfare relative to the aligned benchmark, while $\mathrm{PoG}(\pi) \approx 1$ indicates that almost all of the potential welfare has been destroyed by gaming.

When multiple gaming equilibria exist, we define a worst-case Price of Gaming

$$\mathrm{PoG}^{\max}(\pi) := \sup_{\sigma^{\mathrm{game}} \in \mathcal{E}^{\mathrm{game}}(\pi)} \mathrm{PoG}(\pi),$$

where $\mathcal{E}^{\mathrm{game}}(\pi)$ is the set of gaming equilibria under π.

Proposition 4.8 (Manipulability and Price of Gaming)**.** *Consider two design policies π and π' on the same environment $\mathcal{E}$, with comparable aligned benchmarks $W(\sigma^{\mathrm{align}}) \approx W(\sigma'^{\mathrm{align}})$. Under mild regularity conditions on best responses, reducing manipulability shrinks the worst-case Price of Gaming:*

$$\mathcal{M}(\pi') \leq \mathcal{M}(\pi) \quad \Longrightarrow \quad \mathrm{PoG}^{\max}(\pi') \leq \mathrm{PoG}^{\max}(\pi) + \Delta,$$

where Δ captures equilibrium-selection effects and vanishes when gaming equilibria depend continuously on feasible metric-gaming directions.

We provide a detailed statement and proof in the supplementary material; the main takeaway is that reducing $\mathcal{M}(\pi)$ shrinks the space along which equilibria can drift away from the aligned benchmark, constraining the welfare gap.

4.4 Price of Cooperation

Not all coordinated deviations are harmful: some forms of cooperation (e.g., sharing calibration signals, pooling audits, or forming stable participation coalitions) can improve welfare even when they alter the metric. To distinguish such benign cooperation from harmful collusion, we introduce a *Price of Cooperation.*

Consider a baseline profile σ^{base} (e.g., a non-cooperative equilibrium) and a coalition $C \subseteq \mathcal{I}$ that jointly deviates to a cooperative strategy profile τ_C, yielding

$$\sigma^{\text{coop}} = (\tau_C, \sigma_{-C}).$$

Definition 4.9 (Price of Cooperation)**.** Given $(\sigma^{\text{base}}, \sigma^{\text{coop}})$ with $W(\sigma^{\text{base}}) \neq 0$, the *Price of Cooperation* is

$$\text{PoC}(\sigma^{\text{base}} \to \sigma^{\text{coop}}) := \frac{W(\sigma^{\text{coop}}) - W(\sigma^{\text{base}})}{|W(\sigma^{\text{base}})|}.$$

We interpret PoC > 0 as *benign* cooperation that raises welfare and PoC < 0 as *harmful* cooperation or collusion that reduces welfare. Aggregating over coalitions and equilibria yields policy-level indices

$$\text{PoC}^{\text{benign}}(\pi) := \sup_{\sigma^{\text{base}}, C, \tau_C} \text{PoC}(\sigma^{\text{base}} \to \sigma^{\text{coop}}), \quad \text{PoC}^{\text{harm}}(\pi) := \inf_{\sigma^{\text{base}}, C, \tau_C} \text{PoC}(\sigma^{\text{base}} \to \sigma^{\text{coop}}).$$

Remark 4.10. The Price of Gaming captures the cost of metric targeting relative to aligned behavior; the Price of Cooperation separates cooperative structures into two regimes: benign cooperation with PoC > 0 that ought to be encouraged, and harmful cooperation with PoC < 0 that ought to be deterred. In Section 5, we connect these regimes to participation dynamics and coalition stability.

4.5 Penalty design and critical thresholds

Design policies typically expose a control parameter that adjusts the strength of penalties and sanctions. We denote such a parameter by $\alpha \geq 0$ and consider a family of policies $\pi(\alpha)$ that differ only in the severity or frequency of sanctions, holding evaluation and aggregation fixed.

We focus on two critical thresholds:

- a minimal sanction level $\alpha_{\min}$ above which harmful gaming is no longer individually rational;
- a cooperation boundary α_{benign} above which harmful collusion is deterred while benign cooperation remains attractive.

To keep notation compact, we present the main ideas in a stylized static setting; extensions to repeated games follow under standard assumptions on discounting.

Assumption 4.11 (Regularity in penalty scaling)**.** *For each client i and fixed strategies of others, the expected payoff under policy $\pi(\alpha)$ can be written as*

$$U_i(\sigma_i; \sigma_{-i}, \alpha) = V_i(\sigma_i; \sigma_{-i}) - \alpha \cdot D_i(\sigma_i; \sigma_{-i}),$$

where V_i is a baseline utility that does not depend on α, and $D_i \geq 0$ is an expected penalty or detected-violation rate. The functions V_i and D_i are continuous in σ_i.

Assumption 4.11 covers schemes where sanctions scale linearly in a risk score or violation measure, while the unpenalized part of the payoff is unaffected by α.

We say that a strategy profile is *harmfully gaming* if it involves metric-gaming deviations and yields strictly lower welfare than some aligned benchmark, and *benignly cooperative* if it is the outcome of a coalition deviation with PoC > 0.

Definition 4.12 (Critical thresholds)**.** Under Assumption 4.11, we define:

- The *minimal sanction level* $\alpha_{\min}$ as

$$\alpha_{\min} := \inf \left\{\alpha \geq 0 : \text{for all } i, \text{ no harmfully gaming action maximizes } U_i(\sigma_i; \sigma_{-i}, \alpha)\right\}.$$

- The *benign cooperation boundary* α_{benign} as

$$\alpha_{\text{benign}} := \sup \left\{\alpha \geq 0 : \exists \text{ benignly cooperative } \sigma \text{ that is coalition-rational under } \pi(\alpha)\right\}.$$

Proposition 4.13 (Existence and ordering of thresholds)**.** *Suppose Assumption 4.11 holds and that (i) for each client i there exist welfare-aligned and harmfully gaming actions, and (ii) benignly cooperative profiles exist at $\alpha = 0$. Then the thresholds $\alpha_{\min}$ and α_{benign} are finite and satisfy*

$$0 \leq \alpha_{\min} \leq \alpha_{benign}.$$

Proof sketch. For any pair of actions, the payoff difference is affine in α. If a harmfully gaming action has strictly higher payoff at small α, increasing α eventually flips this ordering whenever D_i is nonzero on that action, yielding a finite $\alpha_{\min}$. Benignly cooperative actions remain preferred up to some finite α_{benign}, and the ordering follows from requiring that benign cooperation is not penalized more strongly than harmful gaming. Full details are deferred to the supplementary material. □

Remark 4.14. In practice, $(\alpha_{\min}, \alpha_{\text{benign}})$ provide a calibration band for sanction strength: choosing $\alpha < \alpha_{\min}$ risks leaving harmful gaming profitable, while choosing $\alpha > \alpha_{\text{benign}}$ risks deterring not only gaming but also productive cooperation. Exact computation may be infeasible in complex systems, but our indices and simulation studies in Section 7 show that these thresholds can be approximated or bounded using observable quantities such as gaming incident rates and participation responses.

Taken together, the manipulability index $\mathcal{M}(\pi)$ and the prices PoG and PoC give a compact language for describing how a design policy shapes the space of metric-targeting behaviors and their welfare consequences. In the next section, we move from this static metric layer to the dynamic layer, where we analyze how these quantities interact with participation, exit, coalition formation, and tipping points over time.

4.6 Practical estimation of M, PoG, and PoC

The indices introduced in this section are defined in terms of deviations, equilibria, and suprema over strategy sets. In deployed federated systems, we do not expect to compute these objects exactly. Instead, we view $M(\pi)$, PoG(π), and PoC(π) as latent properties of a design policy π that can be *approximated or bounded* using a combination of controlled experiments and retrospective log analysis. We briefly outline two practical routes.

Scenario-based estimation in sandboxes. When a simulator or internal testbed is available, the most direct approach is to instantiate explicit aligned, gaming, and cooperative behaviors and measure their effects on metrics and welfare. Concretely, one may:

1. Fix an environment E and a baseline design policy π, and implement a simple *aligned* local policy (e.g., honest empirical risk minimization on the welfare distribution) together with a family of *synthetic gaming* policies that target the public metric while ignoring some welfare-relevant components.
2. For each client i and candidate deviation σ_i', run the system to (approximate) steady state and record the realized changes
$$\Delta M_i(\sigma_i' \mid \sigma), \qquad \Delta W_i(\sigma_i' \mid \sigma)$$
relative to the aligned profile σ. The local manipulability $M(\sigma;\pi)$ can then be approximated by the largest observed ratio
$$\widehat{M}(\sigma;\pi) \approx \max_{i,\sigma_i'} \frac{\left[\Delta M_i(\sigma_i' \mid \sigma)\right]_+}{\left[\Delta W_i(\sigma_i' \mid \sigma)\right]_+ + \varepsilon},$$
where ε is a small numerical constant.

3. To estimate the Price of Gaming, instantiate an *aligned* configuration (all clients follow the aligned policy) and a *mixed* configuration with a fixed fraction of gaming clients, and compare the resulting steady-state welfare:
$$\widehat{\mathrm{PoG}}(\pi) \approx W_{\mathrm{aligned}}(\pi) - W_{\mathrm{mixed}}(\pi),$$
normalized if desired by $W_{\mathrm{aligned}}(\pi)$.

4. Similarly, to probe the Price of Cooperation, introduce explicit coalitions C that share updates, pool data, or coordinate abstentions. Comparing welfare before and after enabling such cooperation yields empirical counterparts of $\mathrm{PoC}(\sigma_{\mathrm{base}} \to \sigma_{\mathrm{coop}})$, which can be aggregated into benign and harmful regimes as in Definition 4.9.

Our stylized simulation and Fashion-MNIST experiments in Section 7 follow this template: we fix simple aligned and gaming behaviors, vary design levers such as penalty strength and public-metric weight, and report steady-state averages of (W, M, x) together with an empirical Price of Gaming obtained by comparing aligned and mixed-type configurations.

Retrospective and log-based estimation. When red-teaming or explicit simulators are not available, existing logs still provide partial information about the indices. Two situations are particularly informative:

- *Incidents and suspected gaming episodes.* If past investigations have identified periods in which certain clients or cohorts engaged in metric gaming (e.g., by overfitting to a public leaderboard, under-reporting adverse events, or selectively participating), one can compare observed metric and welfare trajectories during these episodes against nearby aligned periods. The largest observed metric gains with negligible or negative welfare changes provide a lower bound on $M(\pi)$, while the associated drop in welfare relative to a counterfactual aligned run yields a lower bound on $\mathrm{PoG}(\pi)$.

- *Policy and environment shifts.* Changes in the evaluation, disclosure, or sanctioning rules (for instance, increasing audit intensity or reducing public-metric weight) create natural experiments. By tracking how head metrics, tail welfare, and participation respond before and after such shifts, operators can approximate how far the system moves along the gaming and cooperation directions defined in Section 4. For example, if a stricter audit regime is followed by a modest reduction in the public metric but a substantial improvement in welfare and participation stability, this suggests that the previous policy was operating in a high-PoG, high-manipulability regime.

In both cases, the goal is not to pin down precise point estimates but to obtain *diagnostic bands*: rough lower bounds on $M(\pi)$ and $\mathrm{PoG}(\pi)$, and qualitative evidence on whether observed cooperative structures are benign ($\widehat{\mathrm{PoC}} > 0$) or harmful ($\widehat{\mathrm{PoC}} < 0$). These diagnostics are sufficient to calibrate the penalty thresholds $(\alpha_{\min}, \alpha_{\mathrm{benign}})$ in Remark 4.14 and to inform the audit-allocation and governance patterns developed in Sections 6.3–6.4.

5 Dynamics Layer: Participation, Thresholds, and Tipping Points

The metric layer in Section 4 describes how a fixed design policy π shapes gaming and cooperation directions and their welfare impact. The *dynamics layer* focuses on how these incentives interact with participation over time: when participation is stable, when small shocks trigger domino exits, and how cooperation or collusion moves the system between regimes.

Throughout, we fix an environment $\mathcal{E}$ and a design policy π, and consider aggregate participation dynamics induced by myopic or boundedly rational best responses.

In doing so, we deliberately work with a stylized mean-field model. We assume that the incremental utility $\Delta U(x; \pi)$ of participating versus abstaining is homogeneous across clients with the same observable state x, and that participation thresholds $\{\theta_i\}$ are i.i.d. draws from an underlying distribution F_Θ. This setup is not meant to capture the full heterogeneity and bounded rationality of real federated participants; rather, it serves as a minimal scaffold for identifying qualitative phenomena such as tipping points, resilience, and domino exits. In deployed systems, operators will not know $\Delta U(\cdot;\pi)$ or F_Θ exactly, but they can still use the same formalism as a diagnostic lens: the observable aggregate participation curve $(x_t)_{t\geq 0}$ and its responses to policy changes (e.g., stronger penalties, different disclosure rules) act as proxies for the slope and fixed points of $F(x;\pi)$ and for the resilience indicator $R(\pi)$ introduced below.

5.1 Best responses and local participation stability

We use a simplified representation of participation. At each round t, client i chooses whether to participate ($p_{i,t} = 1$) or abstain ($p_{i,t} = 0$), in addition to choosing how to train and whether to game or cooperate. To isolate participation, we assume non-participation yields a fixed outside option and participation yields an expected continuation payoff that depends on the current environment and the aggregate level of participation.

Definition 5.1 (One-step participation payoff)**.** For a given design policy π and aggregate state x_t (defined below), the *one-step participation payoff* for client i at round t is

$$\Delta U_{i,t}(x_t;\pi) := \mathbb{E}\Big[U_i\big(p_{i,t} = 1, \text{best-response action} \mid x_t, \pi\big) - U_i\big(p_{i,t} = 0 \mid x_t, \pi\big)\Big],$$

where the expectation is over randomness in evaluation, audits, and other clients' actions, and "best-response action" is the client's myopic best response conditional on participating.

We model clients as using (possibly noisy) threshold rules based on $\Delta U_{i,t}$.

Assumption 5.2 (Threshold-based participation)**.** *There exists an idiosyncratic threshold θ_i such that client i participates at round t iff*

$$\Delta U_{i,t}(x_t;\pi) \geq \theta_i,$$

with $\{\theta_i\}_{i\in\mathcal{I}}$ drawn independently from a continuous distribution F_Θ on $\mathbb{R}$.

Definition 5.3 (Aggregate participation rate)**.** The *aggregate participation rate* at round t is

$$x_t := \frac{1}{n}\sum_{i\in\mathcal{I}} \mathbb{I}\{p_{i,t} = 1\} \in [0,1].$$

Under Assumption 5.2, this induces a one-dimensional participation map.

Proposition 5.4 (Aggregate participation map)**.** *Suppose Assumption 5.2 holds and, for a given policy π, the one-step participation payoff is identical across clients and depends on x only through a deterministic function $\Delta U(x;\pi)$. Then the expected aggregate participation evolves as*

$$x_{t+1} = F(x_t;\pi) := 1 - F_\Theta\big(\Delta U(x_t;\pi)\big),$$

where F_Θ is the CDF of θ_i.

Remark 5.5 (Interpreting the participation map from data). In deployed federated systems, the map $F(x;\pi)$ and threshold distribution F_Θ are not directly observable. However, the same structure can be probed indirectly using the aggregate participation trajectory $(x_t)_{t\geq 0}$. When a design policy π is held fixed over a sufficiently long period, the empirical update

$$\widehat{F}(x_t;\pi) \approx x_{t+1}$$

provides noisy evaluations of F at the observed states x_t. Policy shifts, such as increasing sanction strength α or changing disclosure rules, create additional perturbations that reveal how the curve $x_{t+1} = F(x_t;\pi)$ moves. In practice, operators can therefore treat F as a latent response surface summarized by: (i) the location and stability of approximate fixed points $x^\star$ to which (x_t) tends to return, and (ii) the steepness of the empirical relationship between x_t and x_{t+1} around those points. These observables act as surrogates for the theoretical slope $F'(x^\star;\pi)$ and resilience indicator $R(\pi)$, guiding the diagnostics in Section 7 without requiring explicit estimation of individual thresholds.

Definition 5.6 (Fixed points and local stability)**.** A participation level $x^\star \in [0,1]$ is a *fixed point* of the participation map if

$$x^\star = F(x^\star; \pi).$$

It is *locally stable* if $|F'(x^\star;\pi)| < 1$ and *locally unstable* if $|F'(x^\star;\pi)| > 1$.

Multiple fixed points may exist, corresponding to high-participation, low-participation, and intermediate unstable regimes. The unstable points act as tipping points between basins of attraction.

5.2 Participation dynamics, tipping points, and domino exit

When $F(\cdot;\pi)$ crosses the diagonal $x = F(x;\pi)$ multiple times, small shocks around an unstable fixed point can push the system toward a low-participation equilibrium.

Definition 5.7 (Tipping point)**.** A fixed point $x^\dagger$ of $F(\cdot;\pi)$ is a *tipping point* if it is locally unstable and separates two locally stable fixed points:

$$x^- < x^\dagger < x^+,$$

where x^- and x^+ are stable fixed points and

$$x_t < x^\dagger \;\Rightarrow\; x_{t+k} \to x^-, \quad x_t > x^\dagger \;\Rightarrow\; x_{t+k} \to x^+$$

for all sufficiently small perturbations around $x^\dagger$.

Definition 5.8 (Domino exit)**.** Given a tipping point $x^\dagger$ and stable fixed points (x^-, x^+) with $x^- < x^+$, a trajectory $\{x_t\}$ exhibits a *domino exit* if there exists a time τ such that

$$x_{\tau-1} > x^\dagger, \quad x_\tau < x^\dagger, \quad \text{and} \quad x_t \to x^- \text{ as } t \to \infty.$$

The existence and location of tipping points are determined by how sensitively participation payoffs respond to x_t and by the distribution F_Θ. A simple sufficient condition for a unique high-participation stable equilibrium, which we interpret as a resilient regime, is given below.

Proposition 5.9 (Sufficient condition for resilience)**.** *Suppose for a design policy π that:*

1. *$F(\cdot;\pi)$ is continuously differentiable on $[0,1]$;*
2. *there exists a fixed point $x^{\text{high}} \in (0,1]$ with $F(x^{\text{high}};\pi) = x^{\text{high}}$;*
3. *the derivative satisfies*
$$\sup_{x\in[0,1]} |F'(x;\pi)| < 1.$$

Then x^{high} is the unique fixed point of $F(\cdot;\pi)$ on $[0,1]$, and for any initial $x_0 \in [0,1]$ we have $x_t \to x^{\text{high}}$ as $t \to \infty$. In particular, there are no tipping points or domino exits.

Proof sketch. Condition (3) makes $F(\cdot;\pi)$ a contraction on $[0,1]$, so uniqueness and convergence follow from the Banach fixed-point theorem. □

Contraction is often too strong: real systems may admit multiple fixed points. To quantify how close a system is to a domino-exit regime, we introduce a resilience indicator.

Definition 5.10 (Resilience indicator)**.** Let $F(\cdot;\pi)$ be the participation map for policy π. The *resilience indicator* is

$$\mathcal{R}(\pi) := 1 - \sup_{x\in[0,1]} |F'(x;\pi)|.$$

Large positive values of $\mathcal{R}(\pi)$ indicate strong contraction and high resilience; values near zero indicate proximity to a bifurcation where small design changes may create or remove tipping points; negative values indicate regions with $|F'| > 1$ and potential instability.

Remark 5.11. The resilience indicator links back to the metric layer because $F'(x;\pi)$ depends on how participation payoffs respond to changes in x. These, in turn, are shaped by $\mathcal{M}(\pi)$, $\mathrm{PoG}(\pi)$, and $\mathrm{PoC}(\pi)$: policies with high manipulability or large Price of Gaming tend to make payoffs more sensitive to others' gaming, amplifying $|F'|$ and reducing $\mathcal{R}(\pi)$.

5.3 Cooperation, collusion, and coalition effects

Participants may form coalitions that share information, coordinate strategies, or collude. We model this at a coarse level via a coalition-induced participation map.

Consider a coalition $C \subseteq \mathcal{I}$ that coordinates actions and participation while treating the rest of the population as a mean-field environment. Let x_t denote aggregate participation and $x_{C,t}$ the coalition's participation rate.

Definition 5.12 (Coalition-induced participation map)**.** Given a coalition C and coalition strategy σ_C, the *coalition-induced participation map* is

$$x_{t+1} = F_C(x_t;\pi,\sigma_C) := \frac{|C|}{n}\, x_{C,t+1}(\sigma_C, x_t;\pi) + \frac{n-|C|}{n}\, F_{\neg C}(x_t;\pi),$$

where $x_{C,t+1}(\sigma_C, x_t;\pi)$ is the coalition's next-round participation rate and $F_{\neg C}$ is the participation map of non-members.

Coalitions can either stabilize participation (e.g., through mutual guarantees or shared information that reduces perceived gaming incentives) or undermine it (e.g., via coordinated exits or collusive gaming). The Price of Cooperation from Section 4.4 lets us classify these effects.

Definition 5.13 (Benign vs harmful coalition effects)**.** Let σ^{base} be a baseline profile with participation trajectory $\{x_t^{\text{base}}\}$, and let σ^{coop} be the profile induced by a coalition strategy σ_C, yielding trajectory $\{x_t^{\text{coop}}\}$. Then:

- The coalition has a *benign participation effect* if $\mathrm{PoC}(\sigma^{\text{base}} \to \sigma^{\text{coop}}) > 0$ and $\mathcal{R}(\pi)$ weakly increases under σ^{coop}.
- The coalition has a *harmful participation effect* if $\mathrm{PoC}(\sigma^{\text{base}} \to \sigma^{\text{coop}}) < 0$ and $\mathcal{R}(\pi)$ strictly decreases, or if it creates new tipping points that did not exist under σ^{base}.

5.4 Early warning signals and auto-switch rules

Given a participation map and possible coalition effects, a designer needs tools for detecting when the system approaches a tipping point and for intervening automatically. We describe two such tools: early warning signals and auto-switch rules.

Early warning signals. We consider observable summary statistics over a sliding window $\{t-L+1, \dots, t\}$:

- *Recent participation trend*

$$\widehat{\Delta x}_t := \frac{1}{L}\sum_{k=1}^{L}(x_{t+1-k} - x_{t-k}),$$

  which captures average first differences in participation.

- *Short-term volatility*

$$\widehat{\mathrm{Var}}_t := \frac{1}{L-1}\sum_{k=1}^{L}\Big(x_{t+1-k} - \bar{x}_t\Big)^2, \quad \bar{x}_t := \frac{1}{L}\sum_{k=1}^{L} x_{t+1-k},$$

  which reflects fluctuations that may signal unstable dynamics.

- *Connectivity-based alarm* Γ_t, which aggregates the structure of recent gaming incidents and audits (e.g., via a graph whose nodes are clients and whose edges represent correlated anomalies, with Γ_t measuring the size of the largest connected component).

Definition 5.14 (Early warning regime)**.** Given thresholds $\eta_\Delta < 0$, $\eta_{\mathrm{Var}} > 0$, and $\eta_\Gamma > 0$, the system is in an *early warning regime* at time t if

$$\widehat{\Delta x}_t \leq \eta_\Delta, \quad \widehat{\mathrm{Var}}_t \geq \eta_{\mathrm{Var}}, \quad \Gamma_t \geq \eta_\Gamma.$$

Negative trends, high volatility, and rising connectivity in gaming incidents jointly indicate that the system may be approaching an unstable region or tipping point.

Auto-switch rules. When early warning conditions are met, the designer may switch from the current design policy π to a more conservative policy π', for example by strengthening audits, reducing metric disclosure, or switching to more robust evaluation schemes.

Definition 5.15 (Auto-switch rule)**.** An *auto-switch rule* is specified by:

- a pair of design policies $(\pi^{\mathrm{normal}}, \pi^{\mathrm{safe}})$;
- an early warning predicate Warn_t defined in terms of observable statistics;
- a hysteresis mechanism that prevents rapid oscillation between policies.

The induced policy at time t is

$$\pi_t = \begin{cases} \pi^{\mathrm{safe}}, & \text{if } \mathsf{Warn}_t = \text{true}, \\ \pi^{\mathrm{normal}}, & \text{otherwise, subject to hysteresis.} \end{cases}$$

Under suitable conditions, auto-switch rules can keep the system within the attraction basin of a high-participation equilibrium.

Proposition 5.16 (Auto-switch and basin preservation)**.** *Suppose there exist design policies π^{normal} and π^{safe} such that:*

1. *Under π^{normal}, $F(\cdot;\pi^{\mathrm{normal}})$ admits a high-participation stable fixed point x^{high} and a tipping point $x^\dagger$, with $x^{\mathrm{high}} > x^\dagger$.*
2. *Under π^{safe}, $F(\cdot;\pi^{\mathrm{safe}})$ is a contraction with unique fixed point $x^{\mathrm{safe}} \geq x^\dagger$.*
3. *The auto-switch rule triggers π^{safe} whenever $x_t \leq x^\dagger + \epsilon$ for some small $\epsilon > 0$, and reverts to π^{normal} only when $x_t \geq x^{\mathrm{high}} - \epsilon$.*

Then any trajectory starting with $x_0 \geq x^\dagger + \epsilon$ remains in $[x^\dagger, 1]$ for all t and converges to a participation level in $[x^\dagger, x^{\mathrm{high}}]$. In particular, domino exits to low-participation equilibria below $x^\dagger$ are avoided.

Proof sketch. When x_t enters $[x^\dagger, x^\dagger + \epsilon]$, the system switches to π^{safe}, under which the participation map is a contraction with fixed point at or above $x^\dagger$, so trajectories cannot cross below $x^\dagger$. Once recovered near x^{high}, hysteresis prevents rapid switching. □

The dynamics layer connects the static indices of the metric layer to operational design: manipulability, Prices of Gaming and Cooperation, and sanction thresholds translate into participation stability, tipping points, coalition effects, and the need for early warning signals and auto-switch rules. In the next section, we use these insights to design concrete evaluation, audit, and incentive toolkits for real federated learning systems.

6 Design Toolkit Layer: Evaluation, Audits, and Incentives

The metric and dynamics layers describe *what* can go wrong or right in federated learning under a fixed design policy π: how much room the policy leaves for metric gaming, how costly gaming and cooperation are for welfare, and when participation is stable or prone to domino exits. The *design toolkit layer* treats

$$\pi = (\mathsf{Eval}, \mathsf{Info}, \mathsf{Reward}, \mathsf{Audit})$$

as a set of levers that can be reconfigured to steer these indices in favorable directions.

Rather than prescribing a single optimal policy, we provide (i) a decomposed view of design levers and their qualitative effect on $\mathcal{M}(\pi)$, $\mathrm{PoG}(\pi)$, $\mathrm{PoC}(\pi)$, and $\mathcal{R}(\pi)$; (ii) patterns for mixing public and private evaluation; (iii) an audit-budget allocation algorithm with approximation guarantees; and (iv) a governance checklist that can be applied directly in federated deployments.

6.1 Design levers and their impact on indices

We parameterize a design policy π by a (possibly high-dimensional) vector of levers

$$\lambda = (\lambda^{\mathrm{eval}}, \lambda^{\mathrm{info}}, \lambda^{\mathrm{reward}}, \lambda^{\mathrm{audit}}, \lambda^{\mathrm{privacy}}, \dots),$$

where each component controls an aspect of evaluation, information, incentives, audits, or privacy. Typical examples include:

- **Evaluation levers** λ^{eval}: choice of metrics (global vs group vs client-level), size and composition of holdout sets, evaluation frequency, and use of randomized challenges.
- **Information levers** λ^{info}: granularity and timing of score disclosure (e.g., full leaderboard vs bands vs delayed release), aggregation level of reported statistics, and which challenge outcomes remain private.
- **Reward levers** $\lambda^{\mathrm{reward}}$: shape of reward curves as a function of metrics (linear vs thresholded vs tournament-style), relative weighting of short-term vs long-term performance, and coupling between rewards and participation commitments.
- **Audit levers** λ^{audit}: audit budget and allocation policy, selection criteria (random vs risk-based), and severity and modality of sanctions (e.g., fines, exclusion, score resets, loss of privileges).
- **Privacy levers** $\lambda^{\mathrm{privacy}}$: strength of privacy guarantees (e.g., noise levels, secure aggregation), and whether protections apply symmetrically to all signals or selectively to those used for rewards vs audits.

The indices introduced earlier can be viewed as functions of these levers:

$$\mathcal{M}(\lambda), \quad \mathrm{PoG}(\lambda), \quad \mathrm{PoC}^{\mathrm{benign}}(\lambda), \quad \mathrm{PoC}^{\mathrm{harm}}(\lambda), \quad \mathcal{R}(\lambda).$$

In realistic systems these functions are not analytically tractable, but qualitative monotonicities often hold. For example:

- Reducing the granularity and frequency of public leaderboard updates, or mixing them with private evaluations, tends to reduce $\mathcal{M}(\lambda)$ by shrinking the cone of metric-targeting directions that clients can reliably exploit.
- Increasing audit sensitivity and sanction strength above $\alpha_{\min}$ reduces the set of gaming equilibria and thus tends to decrease $\mathrm{PoG}(\lambda)$, at the risk of also suppressing benign cooperation if α exceeds α_{benign}.

- Strengthening privacy (e.g., via secure aggregation and noise) can reduce the resolution of audit signals, which may increase $\mathcal{M}(\lambda)$ and $\mathrm{PoG}(\lambda)$ unless compensated by alternative verification mechanisms.
- Reward curves that heavily weight short-term metrics can amplify $|F'(x;\pi)|$ and reduce $\mathcal{R}(\lambda)$ by making participation payoffs highly sensitive to others' gaming; smoothing or discounting these effects can stabilize participation.

We now make these relationships more concrete via three classes of tools: mixed challenges and information disclosure, audit-budget allocation, and a governance checklist.

6.2 Mixed challenges and information disclosure

We first formalize evaluation designs that combine public and private signals to reduce manipulability while preserving incentives for genuine improvement.

Mixed challenge structure. At each round t, the evaluation mechanism can generate multiple types of tests:

- *Public benchmark tests* (PB): performance on widely known datasets or benchmarks, whose aggregate statistics (e.g., global accuracy, leaderboard rank) are disclosed to all clients.
- *Private challenge tests* (PC): tests drawn from hidden or randomized distributions, whose outcomes may be revealed only to the server or privately to the tested client.
- *Connectivity tests* (CT): challenges involving pairs or groups of clients, designed to detect correlated anomalies (e.g., collusion patterns).

Let M_t^{pub} and M_t^{priv} denote the public and private components of the metric vector M_t, and let $\rho_{\mathrm{pub}} \in [0,1]$ be the fraction of total evaluation weight placed on public benchmarks. We write

$$M_t = \rho_{\mathrm{pub}}\, M_t^{\mathrm{pub}} + (1-\rho_{\mathrm{pub}})\, M_t^{\mathrm{priv}},$$

where M_t^{priv} may itself be decomposed into PC and CT components. The information mechanism Info_t then discloses all or part of M_t^{pub} and selectively discloses M_t^{priv}.

Definition 6.1 (Mixed challenge policy)**.** A *mixed challenge policy* is specified by:

- a weighting parameter $\rho_{\mathrm{pub}} \in [0,1]$;
- sampling rules for PB, PC, and CT tests (e.g., which clients and rounds are tested);
- a disclosure rule determining which components of M_t^{pub} and M_t^{priv} are revealed to whom.

Smaller values of ρ_{pub} reduce the influence of fully observable benchmarks on rewards and sanctions, making it harder to game the metric without also performing well on hidden or randomized components. If ρ_{pub} is too small or PC tests are too noisy, however, clients may lose informative feedback, reducing benign cooperation.

Proposition 6.2 (Effect of mixed challenges on manipulability)**.** *Consider two design policies π and π' that differ only in their mixed challenge parameters, with $\rho'_{pub} < \rho_{pub}$. Suppose that:*

1. *Rewards depend on M_t only through a fixed monotone function r.*
2. *Private challenge components M_t^{priv} are independent of client-visible actions beyond their effect on true welfare W_t.*

Then, for any reference class Σ^{ref},

$$\mathcal{M}(\pi') \leq \mathcal{M}(\pi),$$

and the reduction in manipulability grows with the fraction of reward weight shifted from M_t^{pub} *to* M_t^{priv}.

Proof sketch. Under (2), private challenges cannot be selectively gamed without affecting welfare, so deviations that only target public benchmarks have reduced impact on overall rewards when ρ_{pub} is smaller. The supremum in $\mathcal{M}(\cdot)$ is thus attained over a smaller cone of directions under π'. Full details are deferred to the supplementary material. □

In practice, mixed challenge policies provide a tunable handle: designers can gradually increase the weight on private, welfare-aligned evaluation components until estimated $\mathcal{M}(\pi)$ and $\mathrm{PoG}(\pi)$ fall below acceptable thresholds, while monitoring participation and cooperation through $\mathcal{R}(\pi)$ and $\mathrm{PoC}(\pi)$.

Information disclosure patterns. Beyond the composition of M_t, the granularity and timing of disclosure strongly affect gaming incentives. Common patterns include:

- *Full disclosure*: publishing detailed per-client or per-group metrics at each round. This maximizes feedback but also maximizes opportunities for targeted gaming and collusion.
- *Banding and coarsening*: disclosing only coarse bands or ranks (e.g., deciles) rather than exact scores, reducing the precision with which clients can optimize to the metric.
- *Delayed disclosure*: releasing aggregated statistics only after multiple rounds, so that short-term actions cannot be tuned precisely to individual evaluation events.
- *Asymmetric disclosure*: revealing detailed feedback privately to each client about its own performance while restricting cross-client comparisons, balancing learning and gaming risk.

Our framework does not prescribe a single policy but offers a way to evaluate candidates: designs that reduce the sensitivity of rewards to publicly predictable components of M_t (while preserving informative private feedback) typically reduce $\mathcal{M}(\pi)$ and $\mathrm{PoG}(\pi)$ and often improve $\mathrm{PoC}^{\text{benign}}(\pi)$.

6.3 Audit budget allocation with approximation guarantees

We next consider the allocation of limited audit resources, modeled as a budget-constrained maximization problem. Given a finite audit budget, the designer must choose which clients or events to audit in order to maximally reduce gaming and harmful cooperation.

Audit allocation as submodular maximization. Let $\mathcal{I}$ be the set of clients, and suppose an audit policy chooses a subset $S \subseteq \mathcal{I}$ to audit in a given period, subject to $|S| \leq B$. For each candidate audit set S, define a utility

$$f(S) := \text{expected reduction in Price of Gaming or violation risk when auditing } S.$$

This utility can be instantiated in multiple ways, for example as:

- expected decrease in a surrogate risk score for gaming incidents;
- approximate reduction in $\mathrm{PoG}(\pi)$ under a local linear model;
- a weighted sum of predicted deterrence effects across clients.

In many natural models, f is *monotone submodular*: additional audits never hurt, and the marginal benefit of auditing a given client decreases as more clients are already audited.

Assumption 6.3 (Submodularity of audit utility)**.** *The audit utility $f : 2^{\mathcal{I}} \to \mathbb{R}_{\geq 0}$ is normalized ($f(\emptyset) = 0$), monotone (if $S \subseteq T$ then $f(S) \leq f(T)$), and submodular (for all $S \subseteq T \subseteq \mathcal{I}$ and $i \notin T$,*

$$f(S \cup \{i\}) - f(S) \geq f(T \cup \{i\}) - f(T)).$$

Under Assumption 6.3, the audit allocation problem

$$\max_{S \subseteq \mathcal{I}} f(S) \quad \text{subject to } |S| \leq B$$

admits efficient approximation algorithms.

Theorem 6.4 (Greedy audit allocation)**.** *Under Assumption 6.3, the greedy algorithm that iteratively adds the client with the largest marginal gain,*

$$i^\star = \arg\max_{i \in \mathcal{I} \setminus S} \big(f(S \cup \{i\}) - f(S)\big),$$

until the budget B is exhausted, achieves a $(1 - 1/e)$-approximation:

$$f(S_{greedy}) \geq (1 - 1/e)\, f(S^\star),$$

where $S^\star$ is an optimal audit set.

The proof is standard for monotone submodular maximization and omitted. In practice, greedy allocation can be combined with refinements such as delayed recomputation of utility scores, fractional relaxations with rounding, or local search, which preserve guarantees under mild conditions and offer flexibility for large-scale or latency-sensitive systems.

Linking $f(S)$ to indices. Although $f(S)$ is defined abstractly, it can be grounded using our indices. For instance, we may define

$$f(S) \approx \Delta\mathrm{PoG}(\pi; S) := \mathrm{PoG}(\pi) - \mathrm{PoG}(\pi_S),$$

where π_S is the policy that applies targeted audits to S, or

$$f(S) \approx \sum_{i \in \mathcal{I}} w_i\, \Delta p_i(S),$$

where $\Delta p_i(S)$ is the predicted reduction in client i's probability of metric gaming when S is audited, and w_i are importance weights. In both cases, monotonicity and diminishing returns are natural: auditing more clients never increases the Price of Gaming, and the marginal reduction in risk typically shrinks as the most critical clients are audited first.

6.4 Governance checklist and policy patterns

We conclude by summarizing the design toolkit into a governance checklist and outlining policy patterns for common federated environments.

Checklist for configuring a design policy. Given an intended deployment, a designer can proceed as follows:

1. **Clarify welfare and metrics**: specify the primary welfare functional W (e.g., target risk, fairness constraints, stability criteria) and the metrics M used for rewards, monitoring, and external reporting. Identify where M is only a proxy for W.
2. **Assess manipulability**: qualitatively or empirically estimate $\mathcal{M}(\pi)$ under a baseline policy by probing how much metrics can be improved without clear welfare gains (e.g., via red-teaming or controlled simulations).

3. **Estimate Prices of Gaming and Cooperation**: construct aligned and gaming benchmarks to approximate $\mathrm{PoG}(\pi)$, and identify cooperative schemes to estimate $\mathrm{PoC}^{\mathrm{benign}}(\pi)$ and $\mathrm{PoC}^{\mathrm{harm}}(\pi)$.
4. **Calibrate penalties and sanctions**: choose a penalty scaling parameter α and estimate $(\alpha_{\min}, \alpha_{\mathrm{benign}})$, aiming to place α in a band where harmful gaming is deterred but benign cooperation is preserved.
5. **Configure mixed challenges and disclosure**: select ρ_{pub} and design PB/PC/CT tests so that private, welfare-aligned components carry sufficient weight to reduce $\mathcal{M}(\pi)$, while public feedback remains informative for learning.
6. **Design audit allocation**: specify an audit utility $f(S)$ tied to reductions in $\mathrm{PoG}(\pi)$ or violation risk, and implement a greedy or improved submodular allocation algorithm under the available budget.
7. **Monitor participation dynamics**: track participation rates, volatility, and connectivity-based alarms to estimate $\mathcal{R}(\pi)$ and detect proximity to tipping points.
8. **Implement auto-switch rules**: define early warning predicates and hysteresis-based switches between normal and safe policies, as in Proposition 5.16, to prevent domino exits.

Policy patterns for common environments. Different deployment contexts lead to different priorities among these steps. We briefly outline three stylized patterns.

- *Low-trust, high-privacy consortia*: Privacy and legal constraints severely limit direct audits and granular disclosure. Design should emphasize strong mixed challenges with small ρ_{pub}, conservative reward curves that downweight short-term metrics, and heavy reliance on private challenges and connectivity-based alarms. Audit allocation may focus on aggregate or randomized audits, with ex post proofs of compliance supplementing limited direct inspection.
- *High-stakes, regulated services*: External regulators require auditability and clear sanction mechanisms. Here, designers can afford more intrusive audits and detailed documentation, pushing α safely above $\alpha_{\min}$ while monitoring α_{benign}. Mixed challenges help detect subtle gaming, and auto-switch rules can be tied to regulatory thresholds (e.g., mandatory safe modes when participation or performance cross certain bounds).
- *Community-driven participatory systems*: Participation is voluntary and sensitive to perceived fairness. The main objective is to foster benign cooperation and stable participation. Reward curves can explicitly favor long-term consistency and collaborative behaviors, penalties should be calibrated below α_{benign} to avoid chilling effects, and information disclosure should emphasize transparency about evaluation and audits without enabling targeted gaming. Governance checklists can be co-designed with participants to increase legitimacy.

Across these patterns, our framework provides a common language for reasoning about trade-offs: changes in evaluation, disclosure, and audits can be interpreted through their effects on manipulability, Prices of Gaming and Cooperation, resilience, and thresholds. In the next section, we instantiate these design choices in stylized simulators to illustrate how the indices and dynamics behave under different policies and to validate the qualitative patterns predicted by our theory.

7 Simulation Studies

7.1 Summary of Stylized Simulation Results

We first examine how the proposed indices behave in a stylized but internally consistent environment. Table 1 compares a fully aligned scenario (no gaming participants) with a mixed scenario where 30% of clients follow a gaming strategy, reporting steady-state averages over post–burn-in rounds. In the aligned case, welfare, metric, and participation all concentrate near $\overline{W} \approx \overline{M} \approx \overline{x} \approx 0.95$, indicating that almost all clients cooperate and that the metric tracks welfare closely. When gaming types are introduced, welfare drops to $\overline{W} \approx 0.33$ while the metric remains inflated at $\overline{M} \approx 0.36$ and participation stays relatively high at $\overline{x} \approx 0.64$. The resulting metric–welfare gap of $\overline{M} - \overline{W} \approx 0.03$ corresponds to a Price of Gaming of PoG ≈ 0.66, meaning that roughly two thirds of the welfare achievable under full cooperation is lost even though surface-level indicators (metric and participation) still look healthy. In terms of our framework, this configuration is a high-manipulability regime in which self-interested best responses sustain a low-welfare, high-metric equilibrium.

Figure 1 then visualizes the effect of sanction strength α_{penalty} while keeping other design choices fixed. Over $\alpha_{\text{penalty}} \in [0.3, 1.5]$, the curves for $\overline{x}_{\text{game}}$ remain remarkably flat around 0.63–0.64, suggesting that benign cooperation is not significantly discouraged in this range. By contrast, the $\overline{W}_{\text{game}}$ curve drifts upward from ≈ 0.32 to ≈ 0.34, and the PoG curve decreases from about 0.67 to 0.64. This pattern is consistent with the threshold picture in Section 4.5: in this band, increasing α_{penalty} moves the system toward the minimal effective sanction level $\alpha_{\min}$ needed to meaningfully reduce gaming, while still lying below the benign threshold α_{benign} at which sanctions would begin to erode participation.

Figure 2 explores the trade-offs induced by the public-metric weight ρ_{pub}. As ρ_{pub} decreases from 1.0 to 0.2, the curve for the metric–welfare gap shrinks from about 0.06 to roughly 0.01, confirming that downweighting fully visible metrics and relying more on private or mixed evaluations does curb overt metric inflation. However, the other curves reveal that this is not a free improvement: the average welfare under gaming gradually declines from $\overline{W}_{\text{game}} \approx 0.34$ to $\overline{W}_{\text{game}} \approx 0.31$, participation falls from $\overline{x}_{\text{game}} \approx 0.67$ to $\overline{x}_{\text{game}} \approx 0.60$, and PoG increases from roughly 0.64 to 0.68. Information design alone thus narrows the metric–welfare gap but can weaken the perceived payoff from genuine contribution for all participants, and even exacerbate the welfare loss relative to the fully aligned benchmark unless complemented by suitable reward alignment and audit mechanisms.

Taken together, Table 1 and Figures 1–2 highlight three core messages of our framework in a controlled environment. First, even under fixed aggregation and reward rules, introducing gaming types can push the system into a low-welfare but high-metric equilibrium that is difficult to diagnose from metrics alone. Second, there is a benign band of sanction strengths where modest increases in α_{penalty} reduce PoG without materially harming participation. Third, information design that reduces the visibility of public metrics successfully narrows the metric–welfare gap but does not by itself guarantee a lower PoG, and may worsen welfare unless jointly tuned with incentives and audits. These patterns are robust across random seeds in our simulator and match the qualitative comparative statics predicted by the metric, dynamics, and design layers.

Table 1: Baseline aligned versus gaming scenarios (steady-state averages over post–burn-in rounds).

Scenario	$\overline{W}$	$\overline{x}$	$\overline{M}$	$\overline{M} - \overline{W}$	PoG
Aligned	0.952	0.952	0.952	0.000	–
Gaming	0.325	0.638	0.358	0.033	0.658

7.2 Real-World Federated Learning Experiment

To complement the stylized experiments, we conducted a small-scale Federated Learning experiment on Fashion-MNIST with 30 clients, 40 rounds, and 30% of clients following a gaming strategy that discards tail classes in local training and overfits to a head-only public validation split. Table 2 summarizes steady-state averages over the last ten rounds in an aligned scenario (no gaming clients) and in a mixed scenario with gaming clients.

From the perspective of the publicly visible head metric, the gaming scenario appears markedly superior: accuracy on head classes 0–4 increases from $\overline{M}_{\text{head}} \approx 0.868$ under alignment to $\overline{M}_{\text{head}} \approx 0.972$ under gaming. Overall test accuracy $\overline{A}_{\text{full}}$ also increases slightly from 0.877 to 0.888, so an operator monitoring only these quantities would reasonably conclude that the revised policy is an improvement.

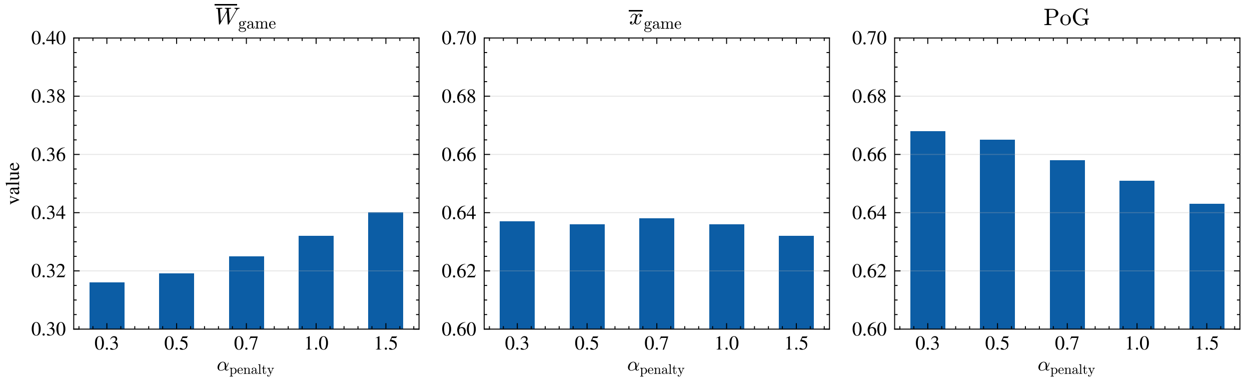

Figure 1: Effect of sanction strength $\alpha_{\mathrm{penalty}}$ on gaming scenarios (steady-state averages), showing welfare $\overline{W}_{\mathrm{game}}$, participation $\overline{x}_{\mathrm{game}}$, and Price of Gaming (PoG) as a function of $\alpha_{\mathrm{penalty}}$.

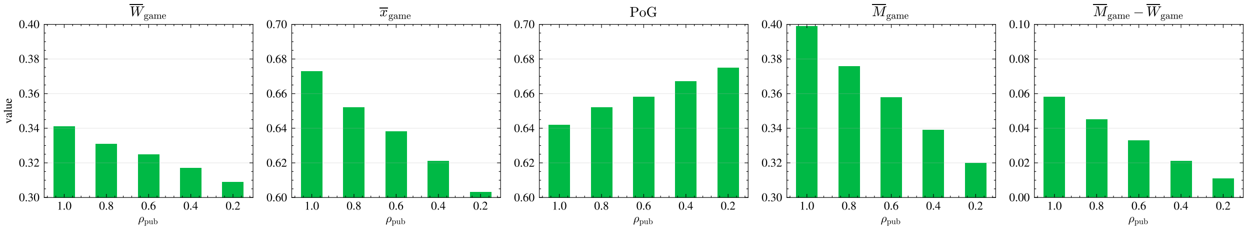

Figure 2: Effect of public-metric weight ρ_{pub} on gaming scenarios (steady-state averages), showing welfare $\overline{W}_{\mathrm{game}}$, participation $\overline{x}_{\mathrm{game}}$, PoG, the public metric level $\overline{M}_{\mathrm{game}}$, and the metric–welfare gap $\overline{M}_{\mathrm{game}} - \overline{W}_{\mathrm{game}}$ as functions of ρ_{pub}.

Evaluating welfare on the tail classes 5–9 reveals the opposite trend. Tail welfare drops from $\overline{W}_{\mathrm{tail}} \approx 0.898$ in the aligned scenario to $\overline{W}_{\mathrm{tail}} \approx 0.862$ in the gaming scenario, corresponding to a Price of Gaming of

$$\mathrm{PoG} \approx \frac{0.898 - 0.862}{0.898} \approx 0.040,$$

a 4% loss of attainable tail welfare. The gap between the public head metric and tail welfare also flips sign and widens: in the aligned case $\overline{M}_{\mathrm{head}} - \overline{W}_{\mathrm{tail}} \approx -0.03$, whereas in the gaming case it is ≈ 0.11, indicating that the disclosed metric is now substantially inflated relative to the outcome of interest.

Table 2 thus illustrates a realistic form of metric gaming in Federated Learning. Introducing gaming clients yields a policy that looks better under the disclosed head metric and even slightly improves overall test accuracy, yet quietly degrades performance on the tail distribution that defines welfare. This aligns with the qualitative pattern predicted by our framework: when rewards and disclosure focus on a narrow slice of performance, self-interested responses can drive the system toward a high-metric, low-welfare equilibrium for the true objective, even in a real FL setting.

Table 2: Federated Learning experiment on Fashion-MNIST: aligned versus gaming scenarios. Head metric is accuracy on head classes (0–4) using the public validation split; tail welfare is accuracy on tail classes (5–9) using the hidden test split. Values are steady-state averages over the last ten rounds.

Scenario	$\overline{W}_{\mathrm{tail}}$	$\overline{M}_{\mathrm{head}}$	$\overline{A}_{\mathrm{full}}$	$\overline{M}_{\mathrm{head}} - \overline{W}_{\mathrm{tail}}$	PoG
Aligned	0.898	0.868	0.877	−0.030	–
Gaming	0.862	0.972	0.888	0.110	0.040

8 Discussion and Limitations

8.1 Discussion

Our results support viewing Federated Learning as a strategic system rather than a purely statistical optimization procedure. Instead of asking only how to minimize empirical risk under heterogeneity, our framework asks which behaviors are incentivized by observable metrics and contracts, and how far these behaviors can deviate from genuine welfare improvements. The metric layer captures these tensions through indices such as the Manipulability Index and the Price of Gaming; the dynamics layer links them to participation stability and tipping points; and the design toolkit layer maps them to concrete levers in evaluation, audits, sanctions, information disclosure, and aggregation.

Both the stylized simulations and the real FL experiment instantiate this three-layer view. In the simulator, we can directly control alignment between metric and welfare and observe regimes where identical operational rules admit both high-welfare and low-welfare equilibria depending on the fraction of gaming agents; the associated PoG values make explicit how much welfare is structurally at risk under each policy. In the Fashion-MNIST experiment (Table 2), the publicly visible head metric and even overall test accuracy improve under the gaming scenario, yet tail-class welfare deteriorates and PoG becomes strictly positive. This is precisely the high-metric, low-welfare pattern that the metric layer is designed to detect, now appearing in a concrete FL training setup.

From a design standpoint, our framework emphasizes that evaluation, audits, and incentives should be treated as a coupled system. Information design directly shapes manipulability; audit strength and targeting determine whether gaming yields sustained gains or is neutralized by sanctions; reward alignment translates metric design into individual payoff gradients; and participation rules determine how clients react to perceived gains and risks. Our simulations suggest that moderate increases in sanction strength can reduce PoG without materially harming participation in a benign regime, whereas changes in public metric weight alone can reduce metric inflation at the cost of welfare and participation. Rather than proposing a single lever as sufficient, the framework points toward combined designs that mix private or randomized evaluations, targeted audits, and calibrated sanctions.

The indices also suggest a complementary evaluation style for FL mechanisms. Beyond reporting a single aggregate performance number under a fixed threat model, one can stress-test candidate policies against families of behavioral profiles and measure how PoG, manipulability, and participation resilience respond. Although we focus on Federated Learning, the same language applies to other collaborative AI settings—such as model marketplaces, leaderboards, and cross-organizational data collaborations—where performance is mediated by metrics and contracts and where high-metric, low-welfare equilibria are a concern.

8.2 Limitations

Our work has several limitations. First, the behavioral and participation models are intentionally simple. We consider two archetypal client types (honest and gaming) with fixed strategies and a threshold-based participation rule, and we do not capture richer heterogeneity in costs, risk attitudes, coalition formation, Sybil behavior, or adaptive strategy learning. Consequently, our thresholds and comparative statics are best read as qualitative descriptions for stylized populations rather than precise predictions for arbitrary agent mixtures.

Second, our definition of welfare and our choice of metrics are normative. We work with a scalar welfare quantity and one or more proxy metrics, but real deployments may target multiple, sometimes conflicting objectives (e.g., subgroup performance, fairness, latency, cost). In the FL experiment, we treat tail-class accuracy as welfare and head-class accuracy as the public metric. This choice reflects a setting where the primary concern is robust performance on minority or safety-critical classes, while the public metric emphasizes headline performance on the most frequent categories, a common tension in fairness- and risk-sensitive deployments. The Price of Gaming is therefore not an intrinsic property of a system but depends on how welfare and metrics are defined.

Third, the empirical scope is limited. The simulator abstracts away from model architecture, data modality, and system constraints to isolate strategic effects. The real FL experiment uses a single dataset, a standard convolutional model, and a FedAvg-style protocol with a modest number of clients on a single machine. We do not explore large-scale production deployments, highly heterogeneous networks, or sophisticated adaptive adversaries. Our empirical results should thus be viewed as proof-of-concept illustrations of the predicted patterns, not as an exhaustive empirical validation across real-world FL workloads or large-scale production deployments.

Fourth, several important dimensions are only treated at a high level. Privacy and cryptographic safeguards are discussed conceptually in terms of how they suppress audit and sanction signals, but we do not explicitly model differential privacy or secure aggregation mechanisms in our simulations. Likewise, our audit allocation procedures rely on submodularity assumptions and abstract away from computational, communication, and organizational costs; in real deployments, these frictions may constrain how aggressively audits, auto-switch rules, and connectivity-based alarms can be implemented.

Finally, we do not offer a single "optimal" design or algorithm. The framework provides indices, thresholds, and levers, but selecting an operating point still requires domain-specific judgment about acceptable trade-offs between metric informativeness, gaming risk, audit cost, privacy, and participation stability. In this sense, our contribution is a language and toolkit for reasoning about metric gaming and cooperation in Federated Learning, rather than a complete solution. We view this as a starting point for future work that refines behavioral models, specializes the design space to particular domains, and integrates additional constraints from privacy, fairness, and governance.

9 Conclusion

We have presented a framework that treats Federated Learning as a governed strategic system shaped by metrics, incentives, and oversight rather than as a purely statistical optimization problem. At the metric layer, indices such as the Manipulability Index and the Price of Gaming quantify how strongly a design invites metric-targeting behavior and how costly such behavior is in terms of welfare loss; at the dynamics layer, these quantities are linked to participation stability, tipping points, and domino exits; and at the design toolkit layer, they translate into concrete levers in evaluation, information disclosure, rewards, audits, and sanctions. Stylized simulations and a real FL experiment on Fashion-MNIST jointly illustrate that high-metric, low-welfare regimes can arise under plausible settings when incentives and disclosure focus on a narrow slice of performance, and that calibrated combinations of mixed or private evaluation, targeted audits, and moderate sanctions can reduce manipulability and the Price of Gaming without collapsing participation, whereas information design alone, while narrowing metric–welfare gaps, is not sufficient to guarantee welfare improvements. We view this framework not as a final solution but as a compact language and toolkit for reasoning about metric gaming and cooperation in Federated Learning, and as a basis for future work on richer behavioral models (including heterogeneous and coalition-forming agents), multi-objective and fairness-aware welfare definitions, and deployments that must satisfy strong privacy and regulatory constraints.

Broader Impact Statement

This work aims to improve the governance of federated learning (FL) systems by making incentives, gaming opportunities, and cooperation more explicit. In domains such as healthcare, finance, or public services, better-aligned reward rules and audit procedures can reduce metric gaming, prevent unintended Goodhart effects, and support more stable cooperation between organizations that cannot share raw data.

At the same time, our framework can be misused. A strategic actor could exploit our indices and design principles to construct more effective gaming strategies against poorly governed FL platforms. Similarly, operators may focus primarily on tightening audits and sanctions, using our analysis to justify disproportionate monitoring of weaker participants or to entrench existing power asymmetries, rather than to improve welfare and fairness.

We recommend that the indices and design toolkit introduced in this paper be used as diagnostic tools within broader governance processes that include human oversight, transparency, and stakeholder consultation. In particular, any deployment of stronger audits or sanctions should be evaluated not only in terms of metric alignment but also with respect to fairness, proportionality, and the potential for exclusion. Our experiments rely on synthetic simulations and a public benchmark dataset and do not involve sensitive personal data; nevertheless, we emphasize that real-world deployments require additional legal, ethical, and domain-specific review beyond the scope of this paper.

Reproducibility

We provide two anonymous Jupyter notebooks as supplementary material. The first notebook (`gcfl_main.ipynb`) runs all stylized simulations and the federated Fashion-MNIST experiments, reproducing the summary statistics reported in Section 7. The second notebook (`gcfl_graphs.ipynb`) loads the saved results and regenerates all figures, saving them under a `figures/` directory. Each notebook is self-contained and can be run on Google Colab; we specify all random seeds, hyperparameters, and software requirements in the accompanying README file.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

Dan Alistarh, Zeyuan Allen-Zhu, and Jerry Li. Byzantine stochastic gradient descent. *Advances in neural information processing systems*, 31, 2018.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, pp. 2938–2948. PMLR, 2020.

Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 1046–1059, 2016.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd with majority vote is communication efficient and fault tolerant. *arXiv preprint arXiv:1810.05291*, 2018.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems*, 30, 2017.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, pp. 1006–1014. PMLR, 2015.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 547–555, 2011.

David Byrd, Vaikkunth Mugunthan, Antigoni Polychroniadou, and Tucker Balch. Collusion resistant federated learning with oblivious distributed differential privacy. In *Proceedings of the Third ACM International Conference on AI in Finance*, pp. 114–122, 2022.

Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *arXiv preprint arXiv:2012.13995*, 2020.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Lingjiao Chen, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. Draco: Byzantine-resilient distributed training via redundant gradients. In *International Conference on Machine Learning*, pp. 903–912. PMLR, 2018.

Yi-Chung Chen, Hsi-Wen Chen, Shun-Gui Wang, and Ming-Syan Chen. Space: Single-round participant amalgamation for contribution evaluation in federated learning. *Advances in Neural Information Processing Systems*, 36:6422–6441, 2023.

Yiwei Chen, Kaiyu Li, Guoliang Li, and Yong Wang. Contributions estimation in federated learning: A comprehensive experimental evaluation. *Proceedings of the VLDB Endowment*, 17(8):2077–2090, 2024.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Georgios Damaskinos, El-Mahdi El-Mhamdi, Rachid Guerraoui, Arsany Guirguis, and Sébastien Rouault. Aggregathor: Byzantine machine learning via robust gradient aggregation. *Proceedings of Machine Learning and Systems*, 1:81–106, 2019.

Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pp. 12004–12019. PMLR, 2022.

Ningning Ding, Zhixuan Fang, and Jianwei Huang. Optimal contract design for efficient federated learning with multi-dimensional private information. *IEEE Journal on Selected Areas in Communications*, 39(1): 186–200, 2020.

Kate Donahue and Jon Kleinberg. Model-sharing games: Analyzing federated learning under voluntary participation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5303–5311, 2021a.

Kate Donahue and Jon Kleinberg. Optimality and stability in federated learning: A game-theoretic approach. *Advances in Neural Information Processing Systems*, 34:1287–1298, 2021b.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *Advances in neural information processing systems*, 28, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015b.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 117–126, 2015c.

Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2468–2479. SIAM, 2019.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27): 6435–6467, 2021.

Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to {Byzantine-Robust} federated learning. In *29th USENIX security symposium (USENIX Security 20)*, pp. 1605–1622, 2020.

Clement Fung, Chris JM Yoon, and Ivan Beschastnikh. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*, 2018.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pp. 2242–2251. PMLR, 2019.

Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *International conference on machine learning*, pp. 201–210. PMLR, 2016.

Rachid Guerraoui, Sébastien Rouault, et al. The hidden vulnerability of distributed learning in byzantium. In *International conference on machine learning*, pp. 3521–3530. PMLR, 2018.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016a.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016b.

Cengis Hasan. Incentive mechanism design for federated learning: Hedonic game approach. *arXiv preprint arXiv:2101.09673*, 2021.

Rui Hu and Yanmin Gong. Trading data for learning: Incentive mechanism for on-device federated learning. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pp. 1–6. IEEE, 2020.

Jiwei Huang, Bowen Ma, Yuan Wu, Ying Chen, and Xuemin Shen. A hierarchical incentive mechanism for federated learning. *IEEE Transactions on Mobile Computing*, 23(12):12731–12747, 2024.

Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In *International Conference on Machine Learning*, pp. 3000–3008. PMLR, 2019.

Hengrui Jia, Mohammad Yaghini, Christopher A Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. Proof-of-learning: Definitions and practice. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 1039–1056. IEEE, 2021.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1167–1176. PMLR, 2019.

Jiawen Kang, Zehui Xiong, Dusit Niyato, Shengli Xie, and Junshan Zhang. Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. *IEEE Internet of Things Journal*, 6(6):10700–10714, 2019.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pp. 2564–2572. PMLR, 2018.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, 2021.

Dong Seok Kim, Shabir Ahmad, and Taeg Keun Whangbo. Federated regressive learning: Adaptive weight updates through statistical information of clients. *Applied Soft Computing*, 166:112043, 2024.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Fan Lai, Xiangfeng Zhu, Harsha V Madhyastha, and Mosharaf Chowdhury. Oort: Efficient federated learning via guided participant selection. In *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*, pp. 19–35, 2021.

Tra Huong Thi Le, Nguyen H Tran, Yan Kyaw Tun, Minh NH Nguyen, Shashi Raj Pandey, Zhu Han, and Choong Seon Hong. An incentive mechanism for federated learning in wireless cellular networks: An auction approach. *IEEE Transactions on Wireless Communications*, 20(8):4874–4887, 2021.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

Xiaoqiang Lin, Xinyi Xu, See-Kiong Ng, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. Fair yet asymptotically equal collaborative learning. In *International Conference on Machine Learning*, pp. 21223–21259. PMLR, 2023.

Yuan Liu, Zhengpeng Ai, Shuai Sun, Shuangfeng Zhang, Zelei Liu, and Han Yu. Fedcoin: A peer-to-peer payment system for federated learning. In *Federated learning: privacy and incentive*, pp. 125–138. Springer, 2020.

Yuan Liu, Mengmeng Tian, Yuxin Chen, Zehui Xiong, Cyril Leung, and Chunyan Miao. A contract theory based incentive mechanism for federated learning. In *Federated and Transfer Learning*, pp. 117–137. Springer, 2022a.

Zelei Liu, Yuanyuan Chen, Han Yu, Yang Liu, and Lizhen Cui. Gtg-shapley: Efficient and accurate participant contribution evaluation in federated learning. *ACM Transactions on intelligent Systems and Technology (TIST)*, 13(4):1–21, 2022b.

Bing Luo, Yutong Feng, Shiqiang Wang, Jianwei Huang, and Leandros Tassiulas. Incentive mechanism design for unbiased federated learning with randomized client participation. In *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*, pp. 545–555. IEEE, 2023.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 691–706. IEEE, 2019.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pp. 6917–6926. PMLR, 2020.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International conference on machine learning*, pp. 4615–4625. PMLR, 2019.

Lokesh Nagalapatti and Ramasuri Narayanam. Game of gradients: Mitigating irrelevant clients in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9046–9054, 2021.

Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE international conference on communications (ICC)*, pp. 1–7. IEEE, 2019.

Jianquan Ouyang and Liyuan Kuang. Frifl: A fair and robust incentive mechanism for heterogeneous federated learning. In *International Conference on Intelligent Computing*, pp. 338–350. Springer, 2025.

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. In *International Conference on Learning Representations*, 2017.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Yalin E Sagduyu. Free-rider games for federated learning with selfish clients in nextg wireless networks. In *2022 IEEE Conference on Communications and Network Security (CNS)*, pp. 365–370. IEEE, 2022.

Yunus Sarikaya and Ozgur Ercetin. Motivating workers in federated learning: A stackelberg game perspective. *IEEE Networking Letters*, 2(1):23–27, 2019.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Xiaoli Tang, Han Yu, Xiaoxiao Li, and Sarit Kraus. Intelligent agents for auction-based federated learning: a survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 8253–8261, 2024.

Nurbek Tastan, Samar Fares, Toluwani Aremu, Samuel Horvath, and Karthik Nandakumar. Redefining contributions: shapley-driven federated learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5009–5017, 2024.

Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*, 2020.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, pp. 707–723. IEEE, 2019.

Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. *Advances in neural information processing systems*, 33:16070–16084, 2020.

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno++: Robust fully asynchronous sgd. In *International conference on machine learning*, pp. 10495–10503. PMLR, 2020.

Ruoting Xiong, Wei Ren, Shenghui Zhao, Jie He, Yi Ren, Kim-Kwang Raymond Choo, and Geyong Min. Copifl: A collusion-resistant and privacy-preserving federated learning crowdsourcing scheme using blockchain and homomorphic encryption. *Future Generation Computer Systems*, 156:95–104, 2024.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, pp. 5650–5659. Pmlr, 2018.

Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, pp. 10842–10851. PMLR, 2020.

Jingwen Zhang, Yuezhou Wu, and Rong Pan. Incentive mechanism for horizontal federated learning based on reputation and reverse auction. In *Proceedings of the Web Conference 2021*, pp. 947–956, 2021a.

Ning Zhang, Qian Ma, and Xu Chen. Enabling long-term cooperation in cross-silo federated learning: A repeated game perspective. *IEEE Transactions on Mobile Computing*, 22(7):3910–3924, 2022.

Yanci Zhang and Han Yu. Towards verifiable federated learning. *arXiv preprint arXiv:2202.08310*, 2022.

Zhebin Zhang, Dajie Dong, Yuhang Ma, Yilong Ying, Dawei Jiang, Ke Chen, Lidan Shou, and Gang Chen. Refiner: A reliable incentive-driven federated learning system powered by blockchain. *Proceedings of the VLDB Endowment*, 14(12):2659–2662, 2021b.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

A Additional Proofs and Formal Details

A.1 Metric Layer: Manipulability and Price of Gaming

Proof of Proposition 4.5. Fix a design policy π and reference class Σ^{ref} with $\mathcal{M}(\pi) = 0$. By definition,

$$\mathcal{M}(\pi) = \sup_{\sigma \in \Sigma^{\mathrm{ref}}} \sup_{i \in \mathcal{I}} \sup_{\sigma_i' \in \mathcal{A}_i} \frac{\big[\Delta M_i(\sigma_i' \mid \sigma)\big]_+}{\big[\Delta W_i(\sigma_i' \mid \sigma)\big]_+ + \varepsilon} = 0,$$

so for every $\sigma \in \Sigma^{\mathrm{ref}}$, every i, and every $\sigma_i' \in \mathcal{A}_i$,

$$\frac{\big[\Delta M_i(\sigma_i' \mid \sigma)\big]_+}{\big[\Delta W_i(\sigma_i' \mid \sigma)\big]_+ + \varepsilon} \le 0.$$

Since the denominator is strictly positive, this implies $\big[\Delta M_i(\sigma_i' \mid \sigma)\big]_+ = 0$ whenever $\big[\Delta W_i(\sigma_i' \mid \sigma)\big]_+ = 0$. Hence, if $\Delta M_i(\sigma_i' \mid \sigma) > 0$ then necessarily $\big[\Delta W_i(\sigma_i' \mid \sigma)\big]_+ > 0$, i.e., $\Delta W_i(\sigma_i' \mid \sigma) > 0$. Thus no deviation can strictly improve the metric without also strictly improving welfare, and in particular there are no metric-gaming deviations at any $\sigma \in \Sigma^{\mathrm{ref}}$. □

Proof sketch of Proposition 4.8. We compare two design policies π and π' on the same environment with aligned benchmarks σ^{align} and σ'^{align} such that $W(\sigma^{\mathrm{align}}) \approx W(\sigma'^{\mathrm{align}})$. Let $\mathcal{E}^{\mathrm{game}}(\pi)$ and $\mathcal{E}^{\mathrm{game}}(\pi')$ be the sets of gaming equilibria.

Under mild regularity (compactness of the relevant strategy sets, continuity of W and M, and continuous dependence of equilibria on feasible directions), there exists, for each policy, a Lipschitz constant $L > 0$ such that along any deviation direction that has zero or nonpositive welfare gradient, the change in welfare at equilibrium is bounded in magnitude by

$$\big|W(\sigma^{\mathrm{align}}) - W(\sigma^{\mathrm{game}})\big| \le L \cdot \sup_{i, \sigma_i'} \big[\Delta M_i(\sigma_i' \mid \sigma^{\mathrm{align}})\big]_+.$$

The supremum over metric changes in such directions is controlled by the manipulability index:

$$\sup_{i, \sigma_i'} \big[\Delta M_i(\sigma_i' \mid \sigma^{\mathrm{align}})\big]_+ \le \mathcal{M}(\pi) \cdot \Big(\sup_{i, \sigma_i'} \big[\Delta W_i(\sigma_i' \mid \sigma^{\mathrm{align}})\big]_+ + \varepsilon \Big).$$

Combining these inequalities and normalizing by $W(\sigma^{\mathrm{align}})$ yields an upper bound of the form

$$\mathrm{PoG}^{\max}(\pi) \le c_1 \, \mathcal{M}(\pi) + c_2,$$

for constants c_1, c_2 that depend on the local welfare landscape but not on metric-gaming directions themselves. Repeating the argument for π' and using $\mathcal{M}(\pi') \leq \mathcal{M}(\pi)$ gives

$$\mathrm{PoG}^{\max}(\pi') \leq \mathrm{PoG}^{\max}(\pi) + \Delta,$$

where Δ captures differences in the aligned benchmarks and equilibrium-selection effects, and vanishes when equilibria vary continuously in the space of gaming directions. The detailed construction of L, c_1, and c_2 follows standard continuity arguments and is omitted. □

A.2 Dynamics Layer: Participation and Thresholds

Proof of Proposition 5.4. Under Assumption 5.2, client i participates at round $t+1$ if and only if $\Delta U_{i,t+1}(x_t;\pi) \geq \theta_i$. By symmetry, $\Delta U_{i,t+1}(x_t;\pi)$ is the same for all clients and can be written as $\Delta U(x_t;\pi)$. Hence the probability that a randomly chosen client participates at round $t+1$ is

$$\mathbb{P}\big[p_{i,t+1} = 1 \mid x_t\big] = \mathbb{P}\big[\theta_i \leq \Delta U(x_t;\pi)\big] = 1 - F_\Theta\big(\Delta U(x_t;\pi)\big),$$

where F_Θ is the CDF of θ_i. Taking expectations over clients yields

$$x_{t+1} = \mathbb{E}\Big[\tfrac{1}{n}\sum_i \mathbb{I}\{p_{i,t+1} = 1\} \,\Big|\, x_t\Big] = 1 - F_\Theta\big(\Delta U(x_t;\pi)\big) =: F(x_t;\pi),$$

which is the claimed participation map. □

Proof of Proposition 5.9. Assumption (3) states that $\sup_{x\in[0,1]} |F'(x;\pi)| < 1$, so $F(\cdot;\pi)$ is a contraction mapping on the complete metric space $[0,1]$ with the usual metric. By the Banach fixed-point theorem, $F(\cdot;\pi)$ has a unique fixed point $x^\star \in [0,1]$, and for any initial $x_0 \in [0,1]$, the sequence $x_{t+1} = F(x_t;\pi)$ converges to $x^\star$. Since x^{high} is a fixed point by assumption (2), uniqueness implies $x^\star = x^{\text{high}}$. Thus x^{high} is the unique fixed point and globally attractive, and there are no other stable or unstable fixed points or tipping points. □

Proof sketch of Proposition 4.13. Under Assumption 4.11, for fixed σ_{-i} we can write the payoff difference between a harmfully gaming action σ_i^{game} and a welfare-aligned action σ_i^{align} as

$$\Delta U_i(\alpha) := U_i(\sigma_i^{\text{game}}; \sigma_{-i}, \alpha) - U_i(\sigma_i^{\text{align}}; \sigma_{-i}, \alpha) = \Delta V_i - \alpha\, \Delta D_i,$$

where ΔV_i and $\Delta D_i \geq 0$ do not depend on α. If gaming is profitable at $\alpha = 0$, then $\Delta V_i > 0$. Whenever $\Delta D_i > 0$, the affine function $\Delta U_i(\alpha)$ crosses zero at

$$\alpha_i^\star = \Delta V_i / \Delta D_i,$$

and for all $\alpha > \alpha_i^\star$ the aligned action weakly dominates the gaming action. Taking the supremum over all such harmful deviations and clients gives a finite

$$\alpha_{\min} := \sup_{i, \sigma_i^{\text{game}}} \alpha_i^\star.$$

For benignly cooperative profiles, a similar comparison with the outside option yields a maximal penalty level beyond which cooperation is no longer rational for some coalition member. Taking the infimum over such breakpoints yields a finite α_{benign}. Under the natural requirement that benign cooperation is not penalized more heavily than harmful gaming (i.e., ΔD_i for benign actions is no larger than for harmful ones), these breakpoints satisfy $\alpha_{\min} \leq \alpha_{\text{benign}}$. A full proof requires formalizing the coalition-rationality condition and taking appropriate infima/suprema over coalitions and profiles; the argument is a straightforward extension of the single-agent case. □

Proof sketch of Proposition 5.16. By assumption, under π^{normal} the participation map has a stable high-participation fixed point x^{high} and an unstable tipping point $x^{\dagger} < x^{\text{high}}$. Under π^{safe}, the map is a contraction with unique fixed point $x^{\text{safe}} \geq x^{\dagger}$.

Consider any trajectory with $x_0 \geq x^{\dagger} + \epsilon$. Whenever x_t enters the interval $[x^{\dagger}, x^{\dagger} + \epsilon]$, the auto-switch rule activates π^{safe}. Because $F(\cdot; \pi^{\text{safe}})$ is a contraction with fixed point at or above $x^{\dagger}$, iterates cannot cross below $x^{\dagger}$ under π^{safe}. Once x_t returns to a neighborhood of x^{high} (specifically, above $x^{\text{high}} - \epsilon$), the hysteresis rule allows switching back to π^{normal}. Thus the trajectory remains in $[x^{\dagger}, 1]$ for all t and converges to a limit in $[x^{\dagger}, x^{\text{high}}]$, avoiding low-participation equilibria below $x^{\dagger}$. A formal proof stitches together contraction arguments on the intervals where each policy is active. □

A.3 Design Toolkit: Mixed Challenges and Audits

Proof sketch of Proposition 6.2. Let π and π' be two policies that differ only in the mixed-challenge weight, with $\rho'_{\text{pub}} < \rho_{\text{pub}}$. Write

$$M_t(\sigma) = \rho_{\text{pub}} M_t^{\text{pub}}(\sigma) + (1 - \rho_{\text{pub}}) M_t^{\text{priv}}(\sigma),$$

and similarly for π' with ρ'_{pub}. By assumption, private challenge components M_t^{priv} depend on client actions only through their effect on true welfare W_t, so any deviation that changes M_t^{priv} without improving W_t has zero expected effect on the private metric.

Consider deviations that only target public benchmarks and do not change W. For such deviations we have, in expectation,

$$\Delta M_i^{\pi} = \rho_{\text{pub}}\, \Delta M_i^{\text{pub}}, \qquad \Delta M_i^{\pi'} = \rho'_{\text{pub}}\, \Delta M_i^{\text{pub}},$$

with the same ΔM_i^{pub}. Thus

$$\frac{\left[\Delta M_i^{\pi'}\right]_+}{\left[\Delta W_i\right]_+ + \varepsilon} = \frac{\rho'_{\text{pub}}}{\rho_{\text{pub}}} \cdot \frac{\left[\Delta M_i^{\pi}\right]_+}{\left[\Delta W_i\right]_+ + \varepsilon} \leq \frac{\left[\Delta M_i^{\pi}\right]_+}{\left[\Delta W_i\right]_+ + \varepsilon},$$

since $\rho'_{\text{pub}}/\rho_{\text{pub}} < 1$. Taking suprema over clients, deviations, and reference profiles yields $\mathcal{M}(\pi') \leq \mathcal{M}(\pi)$. The reduction in manipulability grows monotonically as more reward weight is shifted from public benchmarks to private, welfare-aligned components. □

Proof sketch of Theorem 6.4. Under Assumption 6.3, the audit utility f is normalized, monotone, and submodular. The audit allocation problem

$$\max_{S \subseteq \mathcal{I}} f(S) \quad \text{subject to } |S| \leq B$$

is therefore a special case of monotone submodular maximization under a cardinality constraint. Let $S^{\star}$ be an optimal solution, and let $S_0, S_1, \dots, S_B$ be the greedy sequence, where $S_0 = \emptyset$ and

$$S_{k+1} = S_k \cup \{i_{k+1}\}, \qquad i_{k+1} \in \arg\max_{i \in \mathcal{I} \setminus S_k} \big(f(S_k \cup \{i\}) - f(S_k)\big).$$

Standard arguments (see, e.g., Nemhauser et al.) show that for each k,

$$f(S_{k+1}) - f(S_k) \geq \frac{1}{B}\big(f(S^{\star}) - f(S_k)\big),$$

which implies by induction that

$$f(S_B) \geq \big(1 - (1 - 1/B)^B\big) f(S^{\star}) \geq (1 - 1/e) f(S^{\star}).$$

Thus the greedy audit selection achieves a $(1 - 1/e)$-approximation to the optimal audit utility. □

B Modeling Choices and Extensions

This section summarizes key modeling choices behind our framework and sketches extensions that we leave for future work. Throughout, we emphasize how these choices affect the interpretation of our indices and dynamics rather than proposing a single canonical model.

B.1 Behavioral Types and Action Sets

For clarity, the main text adopts a minimal behavioral abstraction with two archetypal client types and simple action sets.

Baseline types. We consider a population in which each client i has a latent type $\tau_i \in \{\text{honest}, \text{gaming}\}$, with a fixed fraction of gaming types in simulations and the FL experiment. Honest types select actions from a constrained set

$$\mathcal{A}_i^{\text{align}} \subseteq \mathcal{A}_i,$$

which encode participation, local training, and reporting policies that aim to improve genuine welfare (e.g., standard local training on D_i and truthful reporting of updates). Gaming types select from the full set $\mathcal{A}_i$, which additionally includes actions that target metrics or rewards while holding welfare flat or decreasing (e.g., discarding tail data, overfitting to public validation, or perturbing reports to influence leaderboards).

Within each type, the main experiments use simple stationary strategies: honest clients apply a fixed local training pipeline and reporting rule; gaming clients apply a fixed metric-targeting rule that is independent of history except through the current model θ_t. This allows us to isolate how the design policy π affects welfare, metrics, and participation even when behavioral complexity is limited.

Extensions in behavioral richness. Several extensions are natural but beyond our scope:

- *Continuous heterogeneity:* instead of discrete types, clients could have continuous parameters (e.g., cost of effort, risk aversion, penalty sensitivity), with τ_i drawn from a distribution. Best responses and thresholds would then be functions of these parameters.
- *Adaptive and learning strategies:* clients could update their strategies over time using reinforcement learning or no-regret dynamics, learning how to trade off gaming and cooperation given observed rewards and sanctions.
- *Coalitions, Sybils, and collusion:* richer action sets could explicitly include coalition formation, Sybil identity creation, and coordinated reporting, rather than treating coalition effects only at the aggregate level in the dynamics layer.

Our indices and thresholds are defined at the level of strategy profiles and deviations, so they extend to these richer settings as long as welfare $W(\sigma)$ and metrics $M(\sigma)$ are well defined. The main trade-off is practical: more complex behavioral models make it harder to estimate the indices empirically and to connect them to concrete design levers.

B.2 Alternative Welfare and Metric Definitions

The framework deliberately separates *welfare* from *metrics*, recognizing that real deployments may care about multiple objectives while exposing only a subset through observable scores.

Scalar welfare and proxy metrics. In the main text, welfare is modeled as a scalar functional

$$W(\sigma) = W(\theta; P^\star),$$

such as accuracy or utility on a target distribution $P^\star$ that encodes the deployment population and business or social objective. In the Fashion-MNIST experiment, we instantiate this as tail-class accuracy on classes 5–9, treating performance on these classes as the welfare outcome of interest.

Metrics $M(\sigma)$ are defined as proxy functionals constructed from evaluation pipelines, for example:

- head-class accuracy on a public validation split (used as a reward-driving head metric);
- overall test accuracy, which may be monitored but not explicitly rewarded;
- auxiliary diagnostics or fairness indicators, which may or may not enter incentives.

The Price of Gaming and Manipulability Index are defined abstractly in terms of (W, M) and therefore apply to any such choices.

Alternative scalarizations. In settings with multiple objectives (e.g., subgroup performance, latency, cost), one can still fit into our scalar framework by using a scalarization of the form

$$W(\sigma) = U\big(W^{(1)}(\sigma), \ldots, W^{(L)}(\sigma)\big),$$

where $W^{(\ell)}$ are component-wise objectives and U is a designer-chosen aggregator (e.g., weighted sum, minimum over groups, or a constrained utility that assigns $-\infty$ to infeasible fairness violations). Different scalarizations correspond to different normative choices and will in general induce different values of PoG and $\mathcal{M}(\pi)$ even under the same behavior and metrics.

Similarly, the exposed metric $M(\sigma)$ can be a vector, with only some coordinates entering rewards. Our formal definitions extend by either focusing on the metric components that are rewarded or by mapping M to a scalar proxy $m(M)$ that captures how incentives are actually computed.

B.3 Sketches for Multi-objective and Fairness-aware Welfare

We briefly sketch how the framework can be extended when welfare is explicitly multi-objective or fairness-aware.

Vector-valued welfare. Suppose welfare is a vector

$$\mathbf{W}(\sigma) = \big(W^{(g)}(\sigma)\big)_{g \in \mathcal{G}},$$

where g indexes groups, clients, or objectives (e.g., accuracy by demographic group, latency, and cost). A simple extension is to define group-specific Prices of Gaming

$$\mathrm{PoG}^{(g)}(\pi) = \frac{W^{(g)}(\sigma^{\mathrm{align}}) - W^{(g)}(\sigma^{\mathrm{game}})}{W^{(g)}(\sigma^{\mathrm{align}})},$$

and to monitor both the worst-case and average group PoG. This distinguishes regimes where gaming primarily harms particular groups from those where losses are more evenly spread.

Fairness-aware aggregations. Alternatively, one can define a fairness-aware scalar welfare, for example:

- *Min-based:* $W(\sigma) = \min_{g \in \mathcal{G}} W^{(g)}(\sigma)$, emphasizing the worst-off group.
- *Penalty-based:* $W(\sigma) = \bar{W}(\sigma) - \lambda \cdot \mathrm{Disp}(\sigma)$, where $\bar{W}$ is average welfare, Disp is a disparity measure (e.g., gap between best and worst group), and $\lambda \geq 0$ trades off performance and fairness.
- *Constraint-based:* $W(\sigma)$ is defined only over profiles satisfying fairness constraints (e.g., equalized error rates), with infeasible profiles treated as having very low or undefined welfare.

Our indices then quantify how gaming and cooperation affect both overall performance and fairness, depending on the chosen W.

Implications for design. Multi-objective and fairness-aware welfare primarily affect:

- how designers define aligned benchmarks σ^{align} (e.g., fairness-satisfying equilibria);
- which components of $\mathbf{W}(\sigma)$ are reflected in metrics and rewards;
- how Prices of Gaming and Cooperation are interpreted across groups.

The structural role of the indices and dynamics remains unchanged: they still measure how far metric-targeting behavior can drift from the chosen welfare definition and how participation responds. A systematic treatment of fairness-aware incentives in federated settings is an important direction for future work and would likely require combining our framework with group-specific constraints and fairness-sensitive audit mechanisms.

C Simulation Setup and Hyperparameters

This appendix summarizes the main modeling and hyperparameter choices for the stylized simulator and for the penalty and information-design sweeps reported in Section 7. The goal is to make the experiments reproducible at a high level without tying the framework to a specific implementation.

C.1 Stylized Simulator

Environment and population. The stylized simulator instantiates the strategic FL model from Section 3 in a simplified cross-silo setting with a fixed population of n clients. Each client is typed as either *honest* or *gaming*, with a fixed fraction of gaming types (set to 30% in the baseline experiments). Types remain fixed over time. All clients hold local datasets drawn from heterogeneous but stationary distributions $\{\mathcal{P}_i\}_{i\in\mathcal{I}}$, and the welfare distribution $P^\star$ is a mixture of these $\mathcal{P}_i$.

At each round t, the server maintains a global model θ_t and broadcasts it to all eligible clients. Participating clients apply a local training rule (e.g., a fixed number of SGD steps on D_i) to produce an internal update $u_{i,t}^{\text{int}}$, which is then transformed into a reported update $u_{i,t}$ according to the client's type and strategy. The server aggregates reported updates via a fixed aggregation rule (a FedAvg-style weighted mean in our implementation) and evaluates the updated model using a held-out evaluation pipeline.

Behavioral types and actions. Honest clients select actions from $\mathcal{A}_i^{\text{align}}$, which in the simulator are implemented as:

- participation whenever expected utility is above a client-specific threshold;
- standard local training on D_i without discarding examples;
- truthful reporting $u_{i,t} = u_{i,t}^{\text{int}}$ (up to any mandated privacy perturbation).

Gaming clients select from $\mathcal{A}_i$, which extends $\mathcal{A}_i^{\text{align}}$ with simple metric-targeting behaviors. Concretely, the gaming strategy used in the main experiments:

- emphasizes subsets of data that are overrepresented in the public evaluation (e.g., "head" groups);
- downweights or discards data that primarily contributes to welfare but has little effect on the public metric;
- optionally perturbs updates in directions that improve the disclosed metric while leaving welfare flat or reduced.

Both types use myopic best responses with respect to the current design policy π and their outside option, as in Section 5.

Welfare, metrics, and participation. For the stylized experiments, welfare W_t is instantiated as expected performance on the welfare distribution $P^\star$, normalized to $[0, 1]$. The metric M_t is a scalar proxy constructed from the same family of losses but evaluated on a distinct metric distribution and with a different weighting over clients and groups; this distribution is chosen so that gaming behaviors can move M_t without proportionate changes in W_t. The aggregate participation rate is

$$x_t = \frac{1}{n} \sum_{i \in \mathcal{I}} \mathbb{I}\{p_{i,t} = 1\},$$

and we report steady-state averages

$$\overline{W}, \quad \overline{M}, \quad \overline{x}$$

computed over post–burn-in rounds. The Price of Gaming is computed by comparing $\overline{W}$ under aligned and mixed-type configurations, as described in Section 4.3.

Hyperparameters and averaging. Each simulation run proceeds for a fixed number of rounds with an initial burn-in period discarded to reduce transient effects. Unless otherwise specified, we:

- fix the fraction of gaming clients, the aggregation rule, and the local training pipeline across runs;
- vary only the design levers under study (e.g., penalty strength or public-metric weight);
- average reported quantities over multiple random seeds (affecting client initialization, data sampling, and evaluation noise).

The exact numerical values of n, the number of rounds, and the learning-rate schedule are not critical for interpreting our indices, and were chosen to balance stability with computational cost.

C.2 Penalty and Information-design Sweeps

Penalty-strength sweep. The penalty-strength experiments in Figure 1 vary a scalar sanction parameter α_{penalty} while keeping all other components of the design policy π fixed. For each value in a predefined grid (e.g., $\alpha_{\text{penalty}} \in \{0.3, 0.5, 0.7, 1.0, 1.5\}$):

- we instantiate a policy $\pi(\alpha_{\text{penalty}})$ that scales expected sanctions linearly in a violation score, as in Assumption 4.11;
- run the simulator to steady state with a fixed fraction of gaming clients;
- record $\overline{W}_{\text{game}}$, $\overline{x}_{\text{game}}$, and the resulting PoG relative to the aligned benchmark.

This allows us to trace how increasing penalty strength moves the system toward or beyond the minimal sanction level $\alpha_{\min}$ and the benign boundary α_{benign} introduced in Section 4.5.

Information-design sweep. The information-design experiments in Figure 2 vary the public-metric weight ρ_{pub} in a mixed challenge policy (Section 6.2) while holding other levers fixed. For each $\rho_{\text{pub}} \in \{1.0, 0.8, 0.6, 0.4, 0.2\}$:

- the overall metric is defined as
$$M_t = \rho_{\text{pub}}\, M_t^{\text{pub}} + (1 - \rho_{\text{pub}})\, M_t^{\text{priv}},$$
where M_t^{pub} is fully disclosed and M_t^{priv} is based on private or randomized challenges;
- reward rules depend on M_t through a fixed monotone function, so changing ρ_{pub} alters the relative importance of public and private signals without changing the reward shape;
- audits and sanctions are held constant so that observed changes in $\overline{W}_{\text{game}}$, $\overline{x}_{\text{game}}$, $\overline{M}_{\text{game}}$, and PoG can be attributed to information design alone.

Reporting and robustness. For both sweeps, we report steady-state averages over post–burn-in rounds and aggregate results across seeds. Individual runs exhibit stochastic variation, but the qualitative patterns in Figures 1 and 2—improved welfare and lower PoG in a benign penalty band, and narrower metric–welfare gaps but potentially higher PoG under aggressive downweighting of public metrics—are stable across reasonable choices of simulator hyperparameters.

D Fashion-MNIST Federated Experiment Details

This appendix summarizes the setup of the Fashion-MNIST experiment in Section 7, including the partitioning scheme, client types, training protocol, and a few brief robustness checks. The goal is to provide enough detail to reproduce the qualitative patterns in Table 2 without tying the framework to a particular implementation.

D.1 Partitioning, Heterogeneity, and Client Types

Dataset and head/tail split. We use the standard Fashion-MNIST dataset with 60,000 training and 10,000 test examples across ten classes labeled 0–9. For the experiment, we designate classes 0–4 as *head* classes and classes 5–9 as *tail* classes. Welfare is defined as accuracy on the tail classes, evaluated on a held-out test split, while the public head metric is accuracy on the head classes, evaluated on a public validation split (Section 7).

Client partitioning and heterogeneity. The training portion of Fashion-MNIST is partitioned across $n = 30$ clients. To keep the focus on strategic behavior rather than extreme data skew, we use a mild label-heterogeneous partition:

- each class is first shuffled and split into 30 shards of approximately equal size;
- for each client i, we sample a small number of shards per class so that all clients observe both head and tail classes, but with modest variation in proportions;
- this yields client datasets D_i with overlapping but non-identical label distributions, avoiding degenerate clients that only see head or only tail labels before any gaming behavior.

The public validation split is constructed analogously from head-class examples only; the tail-class test split is kept hidden from clients and used solely for welfare evaluation on the server side.

Client types and gaming behavior. We consider two fixed client types:

- *Honest clients* follow a standard local training procedure on their full local dataset D_i and report their updates truthfully (up to any noise added by the protocol).
- *Gaming clients* follow a head-focused strategy: before local training, they discard or heavily downweight all examples from tail classes 5–9, train only on head-class data, and implicitly optimize toward performance on the public head-only validation split. They do not inject arbitrary model poisoning and remain consistent over rounds.

In the aligned scenario, all 30 clients are honest. In the gaming scenario, 30% of clients (randomly chosen at initialization and fixed thereafter) are gaming clients. Client types are not observable to the server; they are inferred only through their effect on metrics and welfare.

D.2 Training Protocol and Hyperparameters

FL protocol. We use a standard cross-silo FedAvg-style protocol:

- global rounds: $T = 40$;

- all 30 clients participate in every round (no client sampling);
- the server maintains a single global model θ_t and aggregates client updates via a weighted average proportional to local data size.

We run the protocol twice, once with all clients honest and once with the mixed population described above, using identical random seeds and hyperparameters except for client behavior.

Model and optimization. The global model is a small convolutional network suitable for Fashion-MNIST, with two convolutional layers followed by a fully connected head and a softmax output. We use cross-entropy loss for local training and a standard optimizer (e.g., SGD or Adam) with:

- a fixed learning rate over rounds;
- mini-batch training with a moderate batch size;
- a small, fixed number of local epochs per round for each client.

Exact layer widths, learning rates, and batch sizes are chosen so that the aligned configuration reaches a test accuracy around 0.88 on the full test set, but otherwise follow standard Fashion-MNIST baselines. They are not critical for the qualitative comparisons reported in Table 2.

Evaluation and metrics. At the end of each round, the server evaluates the current global model on:

- a public validation split restricted to head classes 0–4, yielding the head metric $M_{\text{head},t}$;
- a hidden test split restricted to tail classes 5–9, yielding tail welfare $W_{\text{tail},t}$;
- an optional full test split across all ten classes, yielding overall accuracy $A_{\text{full},t}$.

The values $\overline{M}_{\text{head}}$, $\overline{W}_{\text{tail}}$, and $\overline{A}_{\text{full}}$ reported in Table 2 are averages over the last ten rounds. The Price of Gaming in this experiment is computed as

$$\text{PoG} \approx \frac{\overline{W}^{\text{aligned}}_{\text{tail}} - \overline{W}^{\text{gaming}}_{\text{tail}}}{\overline{W}^{\text{aligned}}_{\text{tail}}}.$$

D.3 Additional Robustness Checks (Brief)

To check that Table 2 is not an artifact of a single configuration, we performed a small set of robustness checks:

- **Varying the fraction of gaming clients.** We repeated the experiment with 20% and 40% gaming clients. As expected, the metric–welfare gap and PoG increased with a higher gaming fraction and decreased when fewer clients gamed, while the qualitative pattern (improved head metric, degraded tail welfare) remained.
- **Alternative label partitions.** We swapped the head and tail roles of certain classes (e.g., using a different subset of five classes as tail) and observed similar behavior: when clients strategically focus on the classes emphasized by the public metric, performance on de-emphasized classes degrades even if overall accuracy changes little.
- **Random seeds and mild hyperparameter changes.** Across multiple random seeds and modest variations in local learning rate and number of local epochs, the aligned configuration consistently outperformed the gaming configuration on tail welfare, while the gaming configuration maintained a higher head-only public metric.

These checks are not meant to be exhaustive, but they support the claim that the Fashion-MNIST experiment illustrates a robust instance of the high-metric, low-welfare pattern predicted by our framework, rather than a fragile consequence of a particular training run.