

From Model Approval to Interface Governance: Federated Learning, Lifecycle Accountability, and AI-Enabled Software as a Medical Device

Liu Ben

Shanghai Jiao Tong University
Shanghai, China
liuben@ieee.org

Abstract

AI-enabled Software as a Medical Device (AI-SaMD) is increasingly governed through lifecycle instruments such as predetermined change control plans and Good Machine Learning Practice. Federated learning has been advanced as a privacy-preserving alternative to centralized training, but it destabilizes a foundational assumption of medical device regulation: that governance can be attached to a discrete artifact. We argue that, under continuous cross-institutional adaptation, federated AI-SaMD cannot be governed as a model alone because responsibility becomes reconstructable only at institutional coordination boundaries. We develop *interface governance* as a necessary accountability layer for those boundaries. We make four contributions. We reconceptualize federated AI-SaMD as distributed clinical infrastructure and specify when an ordinary operational interaction becomes a formal governance interface: when clinically consequential risk, evidence, authority, or contestation moves across institutions and cannot be reconstructed from any single actor's records. We then translate the framework into four minimal implementation conditions: standardized triggers, auditable logs, defined escalation routes, and institution-specific response duties. We also develop a patient-facing pathway for AI-involvement review, disclosure, and institutional response without requiring technical knowledge of federation. Interface governance does not replace existing regulatory mechanisms; it specifies the conditions under which they remain contestable, auditable, and actionable in federated clinical AI.

Introduction

AI-enabled Software as a Medical Device (AI-SaMD) is increasingly governed through lifecycle regulatory instruments such as the U.S. Food and Drug Administration's (FDA) predetermined change control plan (PCCP) (FDA 2024) and joint FDA, Health Canada, and UK Medicines and Healthcare products Regulatory Agency (MHRA) guidance on Good Machine Learning Practice (GMLP) (FDA, Health Canada, and MHRA 2021). These instruments depart from static premarket review by recognizing that machine learning systems drift, that retraining cycles introduce new behaviors, and that post-market surveillance is now an explicit regulatory expectation. Federated learning (FL) complicates this lifecycle in a way that the regulatory frame

has not yet fully absorbed: data, updates, validation, and operational responsibility are distributed across institutions, and the artifact authorized at time t may behave differently across sites at time $t + 1$ because the federation aggregates updates from heterogeneous data nodes.

The problem is not simply that federated AI-SaMD has many actors. Existing accounts of joint accountability already recognize that healthcare AI responsibility is distributed across vendors, hospitals, clinicians, regulators, and sometimes patients. The sharper problem is that responsibility becomes *interface-dependent*. A federation produces clinical behavior through handoffs: local data are summarized for aggregation; updates are validated and deployed; clinicians override or accept recommendations; adverse events are reported through separate channels; patients and sites seek review through complaint, safety, or regulatory pathways. If those handoffs are weakly specified, no participant can reconstruct how a local harm, drift pattern, or contestation claim moved through the system. The governance object is therefore not the model alone but the institutional boundaries through which evidence, authority, obligations, and risk signals travel.

This paper develops *interface governance* as a conceptual and operational accountability layer for federated clinical AI. By an interface we do not mean every ordinary interaction among organizations. We mean a coordination boundary that becomes governance-significant when four conditions are present: (1) the boundary transfers clinically consequential evidence, authority, risk, or contestation; (2) no single actor has full visibility and control over the relevant system behavior; (3) a failure at the boundary would make accountability non-reconstructable from any one actor's records; and (4) the boundary can be stabilized through a trigger, log, escalation route, and named response obligation. These criteria distinguish a formal governance interface from routine coordination such as scheduling a meeting, sending a status update, or exchanging non-clinical administrative information.

The contribution is deliberately modest but practically consequential. Interface governance does not claim to replace PCCP, GMLP, hospital AI committees, adverse-event reporting systems, procurement contracts, or tort and product-liability doctrines. It asks what those mechanisms need in order to work when clinical AI is continuously

adapted across institutions that do not share data, incentives, authority, or legal exposure. The answer is minimal coordination: standardized triggers, auditable logs, defined escalation routes, and institution-specific response obligations at five interfaces—data participation, model update, clinical decision, audit and reporting, and escalation/redress.

The paper makes four contributions. First, it reconceptualizes federated AI-SaMD as distributed clinical infrastructure rather than as a model artifact with external documentation. Second, it specifies when an ordinary operational interaction becomes a formal governance interface: when clinically consequential risk, evidence, authority, or contestation moves across institutions and cannot be reconstructed from any single actor's records. Third, it translates the framework into operational obligations for sponsors, aggregators, hospitals, clinicians, regulators, and procurement agreements: standardized triggers, auditable logs, defined escalation routes, and institution-specific response duties. Fourth, it develops patient-facing contestability and redress in operational terms: how patients know that AI involvement was clinically material, how a challenge is initiated, what information is disclosed, and what institutional responses count as meaningful redress.

Scope. Not all federated learning deployments involve continuous post-market model updating. The framework applies most directly to cross-silo federated SaMD deployments involving periodic or continuous model recalibration mediated through centralized or semi-centralized aggregation infrastructures. It applies most strongly where federated updating, recalibration, or post-deployment adaptation materially affects clinical performance. It applies less directly to static research federations, single-site SaMD, or federated arrangements in which post-deployment adaptation is technically disabled or confined to non-clinical components.

Method and analytic strategy. This paper is neither an empirical deployment study, a systematic review, nor a legal doctrinal analysis. It develops a conceptual governance framework through problem-driven structured synthesis. The framework was derived through iterative comparison across four literatures: healthcare AI governance failures, federated learning risk analysis, SaMD lifecycle regulation, and sociotechnical accountability theory. Literature was selected by relevance to post-deployment governance of distributed clinical AI, with priority given to AIES proceedings, peer-reviewed FL governance studies, and current regulatory guidance. Recurrent governance frictions were grouped along four categories: visibility discontinuities, authority fragmentation, escalation breakdowns, and post-deployment adaptation opacity. Interface obligations were then derived by asking what minimum institutional mechanisms would be required to make a federated SaMD decision reconstructable, challengeable, escalatable, and tied to a response obligation.

The framework is evaluated analytically rather than empirically. It succeeds if it identifies the minimum institutional conditions under which a federated SaMD decision can be reconstructed, challenged, escalated, and assigned a response duty; if it supplies operational responses where PCCP, GMLP, and FL risk matrices remain underspecified;

and if it avoids assigning residual responsibility to clinicians or resource-constrained hospitals that lack the visibility and authority to discharge it.

Background and Related Work

Healthcare AI governance and SaMD lifecycle regulation

AIES scholarship has shifted from abstract AI ethics toward institutional, post-deployment, healthcare-specific governance. Three lines are proximate to this paper. Ganesh, Schiff, and Anderson (2025) describe grassroots oversight in radiology where formal governance has not caught up to deployment realities; this paper asks what cross-institutional structure is required when grassroots local governance is insufficient under federation. Gailmard et al. (2025) propose adverse-event reporting designs for AI; this paper asks which interface conditions make such reporting possible when adverse events are visible only when data, update, override, and audit pathways are connected. Bagave et al. (2025) frame healthcare AI accountability as joint and distributed; this paper asks where, operationally, joint accountability is generated and stabilized in a federated topology. Ganesh (2023) frames responsibility gaps as governance problems; this paper locates responsibility-gap closure at specific institutional boundaries.

On the regulatory side, the FDA's final PCCP guidance permits sponsors to specify modifications in advance over an AI-SaMD lifecycle (FDA 2024). GMLP offers principles for development and monitoring (FDA, Health Canada, and MHRA 2021). IMDRF and the EU AI Act establish lifecycle and post-market obligations relevant to medical AI (International Medical Device Regulators Forum 2017; European Union 2024). Recent FL-specific governance proposals offer federation risk matrices (Bottomly et al. 2026). These instruments remain essential, but they tend to treat governance as planned modification, documentation, and monitoring. Federated learning places pressure on three assumptions that often remain implicit: that change can be bounded in advance, that the sponsor can observe the clinically relevant consequences of change, and that adverse events can be attributed through records held by a single institution. Interface governance supplements these instruments by specifying the cross-institutional handoffs that must be logged, escalated, and assigned a response duty.

Federated learning in healthcare

In this paper, *federated AI* means cross-silo clinical machine learning in which hospitals or other institutions retain raw patient data locally, exchange model updates or sufficient statistics through a coordinating sponsor or aggregator, and deploy a shared or related AI-SaMD whose post-deployment behavior may change through periodic or continuous federation. The term therefore excludes ordinary multi-site procurement of a static model and research-only federations that cannot affect clinical use.

FL supports multi-institutional learning without raw data exchange (Rieke et al. 2020; Sheller et al. 2020), but the literature documents persistent risks: gradient leakage (Zhu,

Liu, and Han 2019), poisoning by adversarial nodes (Bagdasaryan et al. 2020), statistical heterogeneity that disadvantages minority sites (Li et al. 2020), and unequal institutional capacity to participate in governance. FL decentralizes data storage, but it does not equalize the ability to define data quality, contest model behavior, or shape update governance.

Federated adaptation is not merely organizational decentralization. Aggregation under non-IID data distributions, asynchronous update cycles, and uneven site participation can produce locally divergent performance trajectories even when global validation metrics remain stable (Li et al. 2020). Model aggregation compresses heterogeneous local updates into a global parameter state whose downstream clinical implications may remain site-specific and only partially observable. A federation can satisfy aggregate validation thresholds while degrading subgroup- or site-specific clinical performance. The governance problem is therefore not simply privacy-preserving learning; it is clinically consequential adaptation across statistically and institutionally heterogeneous sites.

Sociotechnical accountability

Three lines of critical theory ground the framework. Infrastructure studies (Star and Ruhleder 1996; Bowker and Star 1999; Edwards 2010), institutional theory (Abbott 1988; March and Olsen 1989; Scott 2014), and Galison's (1997) account of trading zones supply the vocabulary for coordination without convergence: heterogeneous institutions can interoperate when boundary protocols stabilize exchange. Sociotechnical critique (Suchman 2007; Selbst et al. 2019) and Elish's (2019) moral crumple zone show how human operators in automated systems can be positioned as final safeguards while lacking the visibility and authority to discharge that role. Work on accountability and contestability (Nissenbaum 1996; Ananny and Crawford 2018; Edwards and Veale 2017; Kaminski 2019), the principles-to-mechanisms literature (Mittelstadt 2019; Floridi et al. 2018; Jobin, Ienca, and Vayena 2019; Schiff et al. 2021), and the responsibility-gap tradition (Matthias 2004; Santoni de Sio and Mecacci 2021) together motivate reconstructability rather than total transparency.

Three terms are used in a narrow sense. *Interface-dependent accountability* means that responsibility can be reconstructed only at the points where institutions hand off evidence, authority, or response duties. *Contestable traceability* means that a patient, clinician, site, sponsor, or regulator can reconstruct enough of the relevant pathway to challenge and escalate a decision, without requiring total visibility into proprietary models or patient-level records. *Residual governance* means the accountability layer that remains operationally necessary when centralized control, total transparency, and normative consensus are unavailable.

From Deployment Frictions to Governance Interfaces

Existing healthcare AI deployments already exhibit fragmented override practices, inconsistent reporting pathways, and institution-specific governance improvisation (Ganesh,

Schiff, and Anderson 2025; Gailmard et al. 2025). Federation intensifies these tensions by distributing consequential updates across sites that lack unified visibility or coordinated escalation mechanisms. Five frictions are recurrent.

First, override decisions are recorded inconsistently or not at all across hospitals. Second, post-market reporting lags may reflect reporting-channel fragmentation rather than absence of harm. Third, site-level data heterogeneity may be invisible at the federation level even when it materially affects local performance. Fourth, federated update cycles may propagate consequential changes without site-level revalidation. Fifth, governance authority is dispersed across vendor, aggregator, hospital, clinician, and regulator without a defined adjudication procedure. None of these frictions is unique to federation; federation aggregates them into a single accountability problem.

Dominant governance alternatives become insufficient under federation for structural reasons. Ethical principles remain governance-indeterminate unless indexed to institutions, triggers, escalation routes, and response obligations. Total transparency is impossible and often undesirable: sites cannot expose patient-level data, vendors may not expose proprietary procedures, and regulators cannot consume unbounded volumes of operational information. Human-in-the-loop oversight can convert clinicians into moral crumple zones if they are asked to absorb responsibility for update dynamics they cannot see or control. Centralized regulation cannot alone govern a federation in which clinically relevant behavior emerges across distributed sites. Normative consensus cannot be assumed among hospitals, sponsors, clinicians, regulators, and patient communities with divergent incentives, capacities, and legal exposure.

The right response is not more abstract principle, more data visibility, or more formal consensus. It is to specify the boundary mechanisms through which disagreement, partial visibility, and fragmented authority remain governable. Table 1 identifies those mechanisms at the level of deployment frictions rather than at the level of general values.

Interface Governance

Definition and threshold criteria

Interface governance begins from a different ontology of the governance object. Governance occurs at institutional boundaries. Coordination is intentionally minimal rather than maximal. The interface, not the model alone, is the operative unit of accountability.

In federated clinical AI, accountability is generated relationally rather than locally: no single institution possesses sufficient visibility, authority, and operational control to independently govern system behavior. The interface is the locus at which relational accountability becomes operationally tractable. Existing accountability models often specify who is responsible; interface governance specifies where accountability becomes reconstructable under federation.

Not every coordination point is a governance interface. A coordination point becomes a governance interface when it satisfies the four threshold criteria in Table 2. These criteria are meant to avoid concept inflation. Routine coordination

Observed friction	Literature basis	Why artifact-centered governance misses it	Derived interface requirement
Local override inconsistency across hospitals	(Ganesh, Schiff, and Anderson 2025)	The model record does not capture how clinicians at different sites accepted, questioned, or overrode outputs.	Clinical decision interface with override logging, rationale categories, and local drift surfacing.
Adverse-event underreporting in healthcare AI	(Gailmard et al. 2025)	Absence of reports may reflect fragmented channels, not absence of harm.	Audit and reporting interface with trigger-bound escalation and reporting-lag monitoring.
FL site heterogeneity and non-IID dynamics	(Li et al. 2020; Eden et al. 2025)	Aggregate validation can obscure local performance degradation.	Data participation interface with site-specific quality, representativeness, and contribution metadata.
Joint accountability without operational addressing	(Bagave et al. 2025; Ganesh 2023)	Naming multiple responsible actors does not specify who must respond when a boundary fails.	Interface ownership with institution-specific response duties.
PCCP boundedness assumption under continuous adaptation	(FDA 2024; Bottomly et al. 2026)	Planned modifications may not capture emergent site-specific consequences of federated updates.	Model update interface with update provenance, local performance gates, and rollback criteria.

Table 1: From deployment frictions to interface requirements

involves information exchange that remains reconstructable within ordinary institutional records. A governance interface involves a handoff whose failure would make clinical responsibility, evidence, or redress non-reconstructable unless the boundary is separately governed.

Five interfaces from four requirements

The five interfaces are derived from four persistent governance requirements: participation, adaptation, contestation, and attribution. Contestation appears in two registers: epistemic contestation by clinicians at the decision point and institutional contestation by patients, sites, or regulators after a challenge or harm signal. Table 3 names the interfaces and their governing questions.

Minimal coordination obligations

Attempts at maximal governance under federation produce visibility overload, institutional deadlock, and procedural noncompliance generated by incompatible authority structures. Minimal coordination is therefore not a weak form of governance; it is a design constraint for federated institutions. We define minimal coordination as the smallest set of interoperable obligations necessary to complete an accountability chain from detection to evidence, transmission, and action.

Minimal coordination becomes a mechanism only when each interface discharges four operational obligations. Standardized triggers make a deviation reviewable; auditable logs preserve the evidence needed for reconstruction; defined escalation routes transmit the evidence to an actor with authority; and institution-specific response obligations close the loop by requiring action. The set is minimal by a counter-

factual necessity test. Remove the trigger and no rule moves the system from ordinary operation to review. Remove the log and the event cannot be reconstructed. Remove the route and the signal remains stranded. Remove the response duty and a recipient may acknowledge the signal without acting. The four obligations are jointly sufficient for a procedural accountability chain, although not for substantive clinical safety, which still depends on device- and workflow-specific standards.

For implementation, each obligation must be complete enough to test. A trigger specification names the metric or event, reference population or baseline, observation window, uncertainty rule, crossing condition, and immediate action. A log identifies the device and model version, update provenance, site and workflow, affected subgroup where relevant, initiating signal, timestamps, decision rationale, response owner, and action taken. An escalation route identifies the sender, authorized recipient, transmission channel, acknowledgement and response windows, and fallback recipient if the first route fails. The response duty identifies the actor, permissible actions, deadline, and closure evidence. Numeric thresholds remain device-specific, but these fields are not optional.

Contestable traceability. Interface governance reorganizes the accountability question. Transparency asks who can see. Contestable traceability asks who can reconstruct, who can challenge, who can escalate, and who must respond. It requires only that institutionally consequential actions can be reconstructed, challenged, and escalated through defined procedural pathways. Federation transforms accountability from a visibility problem into a reconstructability problem. The result is not a fully transparent federation but a con-

Criterion	Routine coordination	Governance interface
Clinical materiality	The interaction has no direct effect on clinical risk, update validity, or patient contestability.	The interaction transfers evidence, authority, risk, or contestation that can materially affect clinical use or post-market obligations.
Fragmented visibility or authority	One actor can reconstruct the relevant decision path from ordinary records.	No single actor can reconstruct the decision path without boundary-level records from other institutions.
Accountability failure mode	Failure creates delay, inefficiency, or ordinary coordination error.	Failure makes responsibility, escalation, or patient redress non-reconstructable.
Governability	Additional governance would add documentation without improving accountability.	A trigger, log, escalation route, and named response obligation can materially improve accountability.

Table 2: Criteria distinguishing routine coordination from a governance interface

Requirement	Interface	Function	Governance question
Participation	Data participation	Controls local data contribution, quality, representativeness, and participation status.	Who participates, under what quality and representativeness conditions, and with what support?
Adaptation	Model update	Validates, propagates, withholds, or rolls back updates.	When does an update require site-level review, logging, rollback, or reapproval?
Contestation (epistemic)	Clinical decision	Mediates AI output, professional judgment, disclosure, and override practice.	How can clinicians understand, override, document, and dispute AI outputs without becoming residual liability sinks?
Attribution	Audit and reporting	Records risk evidence across the federation.	What adverse events, drift signals, overrides, and reporting lags must be logged and surfaced?
Contestation (institutional)	Escalation and redress	Handles formal challenge, harm review, corrective action, and patient-facing response.	Who can challenge the system, through what route, and what response is triggered?

Table 3: Five governance interfaces

testable one.

Patient asymmetry is the limit case. For most patients, the federation itself is institutionally invisible. The pathways through which data, updates, and risk signals travel remain inaccessible even when AI materially shapes care outcomes. Patient rights therefore cannot depend on understanding the model, the federation, or the update history. They must depend on accessible routes to contestation and defined institutional responses.

From Framework to Implementation

The main practical challenge is to translate interface governance into obligations that can be adopted without creating an entirely new regulatory institution. This section specifies four implementation channels: regulatory lifecycle instruments, sponsor and aggregator responsibilities, hospital governance and procurement, and patient-facing redress.

Regulatory translation

PCCP annex. For federated SaMD, a PCCP could include a federation-interface annex. The annex would not preapprove every future update. It would specify update cate-

gories, site-specific monitoring metrics, local performance gates, rollback conditions, and the evidence required when an update degrades performance at a minority site while improving aggregate performance. A PCCP would therefore remain a planned-change instrument, but the plan would include boundary conditions for unplanned or asymmetric consequences of federated adaptation.

Sponsor and aggregator duties. Sponsors and aggregators should be responsible for update provenance, versioning, site-level performance dashboards, and escalation infrastructure. They are the actors most likely to observe cross-site patterns. Their duty is not to assume all hospital-level clinical responsibility, but to make site-level consequences visible enough for hospitals, clinicians, and regulators to act.

Hospital governance. Hospitals should locate federated SaMD review within existing patient safety, quality improvement, radiology governance, or AI oversight committees. The hospital duty is to maintain local override records, report site-specific drift or complaints, and decide whether local deployment should be paused, restricted, or accompanied by additional review.

Post-market surveillance. Regulators and standards bod-

Panel A. Triggers and evidence artifacts

Interface	Trigger	Evidence artifact
Data participation	Quality, missingness, coding, label provenance, representativeness, or demographic-shift signal crosses a pre-specified control limit.	Contribution metadata; data-quality report; site population summary; preprocessing and label-provenance record.
Model update	Candidate update improves aggregate performance but worsens site, subgroup, calibration, or OOD performance beyond a pre-specified bound.	Update provenance; model version; validation report; local performance dashboard; rollback comparison.
Clinical decision	Override rate, disagreement pattern, alert fatigue, or clinician complaint exceeds historical baseline or control-limit threshold.	AI output; model and update version; clinician override or acceptance; rationale category; disclosure status.
Audit and reporting	Adverse event, near miss, repeated override cluster, reporting lag, or unfiled-event pattern appears.	Event report; near-miss record; override aggregate; response timestamp; cross-site comparison.
Escalation and redress	Patient, clinician, site, or regulator files a challenge; AI involvement is alleged or clinically material.	Challenge log; AI-use disclosure; relevant model/update version; case review record; outcome notice.

Panel B. Ownership and response duties

Interface	Primary response owner	Escalation route	Minimum response
Data participation	Site data steward with aggregator support.	Federation data review board or sponsor safety function.	Pause or qualify site contribution; provide remediation support; document whether affected outputs require review.
Model update	Sponsor or aggregator.	Sponsor review board; hospital AI committee; regulator when reporting threshold is met.	Withhold, stage, or roll back update; conduct site-level revalidation; document relation to PCCP scope.
Clinical decision	Clinician workflow lead and hospital safety officer.	Hospital safety committee; sponsor if pattern suggests model drift.	Surface drift warning; review affected cases; adjust workflow or pause local use under hospital safety policy.
Audit and reporting	Hospital safety office and aggregator.	Regulator post-market surveillance or sponsor complaint handling system.	File site report within applicable reporting window; reconcile records across hospital, sponsor, and aggregator.
Escalation and redress	Hospital complaint office with sponsor and regulator cooperation where needed.	Independent review, regulator, or external adjudication pathway when local review is insufficient.	Acknowledge filing; determine material AI involvement; conduct secondary review; provide outcome, correction, or escalation record.

Table 4: Operational matrix across interfaces

ies can specify minimum reporting fields: device and model version, update provenance, site-level performance signal, affected subgroup or workflow, override pattern, complaint or adverse-event status, response owner, and action taken. Interface governance therefore strengthens post-market surveillance by making reporting routes and evidence artifacts specific enough to be actionable.

Procurement and contracts. Procurement agreements are the practical bridge between conceptual framework and local adoption. Contracts can require audit-log access, site-level dashboards, update notices, response windows, rollback rights, complaint-routing cooperation, and support for lower-capacity sites.

Liability and contract interaction. Interface obligations do not displace tort, product-liability, or contract doctrine; they change what those doctrines have to work with. Au-

ditable logs and response records supply the evidentiary record that malpractice and product-liability proceedings currently lack: which model version acted, which trigger fired, who was notified, and whether the named owner responded within the agreed window. Procurement terms that encode response duties function as ex ante contractual risk allocation among sponsor, aggregator, and hospitals, giving courts and insurers documented duty assignments rather than post hoc reconstruction. Documented triggers and response windows can likewise inform the applicable standard of care: a clinician who escalates through a defined route holds evidence of a discharged duty, narrowing the moral crumple zone (Elish 2019), while an unmet response duty is attributable to its named owner. Which doctrinal routes absorb these records—negligence, warranty, or regulatory enforcement—remains jurisdiction-specific and is a ques-

tion for legal scholarship rather than this framework.

Operational matrix

Table 4 instantiates the framework as an implementation matrix. The examples are not universal clinical thresholds; they are categories that should be specified in the PCCP, federation agreement, procurement contract, or hospital safety policy for a given device and workflow. For example, a model-update gate is not simply “performance declined.” It identifies the metric and subgroup, the local baseline and evaluation window, the uncertainty-adjusted crossing rule, who receives the alert, and whether crossing the rule causes staged deployment, withholding, or rollback.

Patient-facing contestability and redress

Patient-facing contestability is underdeveloped if it merely says that patients should be able to challenge AI. It must specify how they learn that AI involvement may matter, how they initiate a challenge, what information is disclosed, and what redress can mean.

How patients know AI involvement was clinically material. Patients do not need a technical explanation of federation. They need a clinically meaningful notice when AI-SaMD materially contributed to a diagnostic, triage, or treatment pathway. In practice, this can occur through the radiology report, discharge summary, after-visit summary, patient portal, or informed-use notice. The notice should identify that AI-assisted analysis was used, that a clinician reviewed the output, and that patients may request review if they believe the result affected care.

How a challenge is initiated. A patient should be able to initiate review through existing hospital complaint, patient safety, medical records, or patient portal channels. The patient should not be required to identify the model architecture, the federation, or the relevant update. A challenge should be valid if the patient alleges that an AI-assisted result materially affected diagnosis, triage, treatment, or delay in care.

What information is disclosed. Disclosure should be contestability-oriented rather than transparency-maximalist. A patient should receive enough information to understand whether AI was materially involved, what clinical decision or report it affected, whether a clinician accepted or overrode the output, whether the case fell within a known drift or update concern, and what review occurred. Patient-level records and proprietary model details need not be exposed beyond what is necessary for challenge and review.

What redress involves. Meaningful redress can include secondary clinical review, correction or addendum to a report, notice to the treating clinician, reclassification of the event as reportable, explanation of why AI involvement was or was not material, local workflow change, update rollback or pause, and escalation to regulator or independent review when the hospital cannot resolve the challenge. Redress is therefore not limited to compensation or apology; it is the defined institutional response that makes the challenge consequential.

Capacity-sensitive implementation

Federated governance can reproduce institutional inequality if lower-capacity hospitals are given the same nominal obligations without the same infrastructure. Interface governance therefore requires scaled, not merely symmetric, obligations.

Three design principles follow. First, the federation coordinator or sponsor should provide shared monitoring infrastructure, standardized reporting templates, and technical support for site-level performance review. Second, lower-capacity sites should be allowed to discharge some obligations through shared services, regional data stewards, or regulator-supported templates rather than building full internal AI governance units. Third, accountability should not be decoupled by treating under-resourced sites as the source of risk while giving them the least power to define thresholds, contest updates, or pause deployment. Capacity-sensitive implementation makes interface obligations feasible without making them hollow.

Federated Radiology Scenario

The following analytic case is stylized but grounded in documented deployment frictions in radiology governance, federated learning implementation, and post-market AI monitoring (Ganesh, Schiff, and Anderson 2025; Bottomly et al. 2026; Eden et al. 2025; Gailmard et al. 2025). A federated radiology AI-SaMD is deployed across five hospitals. One site serves a demographically distinct population underrepresented in the federation. Over six months, the global model’s aggregate performance improves, and the sponsor’s PCCP-specified aggregate validation thresholds are met. At the underrepresented site, however, local sensitivity for a clinically important finding declines below the site’s historical baseline; the override rate rises above its local control limit; and delayed adverse-event reports begin to appear after clinicians discover missed cases through downstream follow-up.

Without interface governance, four failures unfold in parallel. Local drift remains hidden because aggregate performance dominates update validation. Override patterns remain institutionally isolated because clinicians document disagreement in local workflow notes rather than in a federation-visible log. PCCP thresholds remain formally satisfied because the planned-change protocol was written around aggregate validation rather than asymmetric local degradation. Accountability collapses into blame diffusion: the vendor points to hospital configuration, the hospital points to training distribution, clinicians point to workflow pressure, and the regulator sees no unified reportable signal.

Under interface governance, the same facts trigger a different sequence. The data participation interface flags the demographic asymmetry as a representativeness and contribution-quality concern. The model update interface routes the candidate update to site-level review because local performance fell below the pre-specified bound even though aggregate performance improved. The clinical decision interface logs rising override rates with rationale categories and surfaces a site-level drift warning to radiolo-

gists. The audit and reporting interface reconciles delayed adverse-event reports with override clusters, making reporting lag itself a governance signal. The escalation and redress interface gives the affected site, clinicians, and patients a defined route to request secondary review, obtain an AI-involvement determination, and trigger rollback, workflow change, or regulator notice where appropriate.

The scenario illustrates why the model is not the only locus of governance. The failure arises from continuous cross-institutional adaptation that incrementally reshapes local clinical performance. The corrective response is not merely to inspect the model. It is to ensure that the boundaries through which data, updates, clinical judgment, audit evidence, and challenges move are reconstructable and assigned response duties.

Discussion and Conclusion

Interface governance does not replace technical robustness, safety engineering, formal regulation, hospital oversight, or legal accountability. It specifies the accountability layer that becomes necessary once distributed infrastructures exceed the governance capacity of artifact-centered oversight alone.

The framework is intentionally residual. It applies where centralization, total transparency, and shared normative consensus are unavailable or undesirable. Healthcare reveals the interface problem with unusual clarity because clinical authority is separated from model ownership, liability from technical maintenance, patient risk from update control, and local clinical judgment from federation-level behavior. These separations make the boundary between institutions a primary site of accountability.

The framework also has limits. It is not a regulatory instrument by itself; it depends on standards bodies, regulators, sponsors, and procurement processes to translate it into binding obligations. It does not specify substantive clinical thresholds for every device or workflow. It does not resolve ultimate tort, product-liability, contract, or professional-liability questions. It does not prevent capture or eliminate institutional capacity asymmetry. What it does specify is the minimum accountability architecture necessary for contestable coordination under federation.

Future research should test interface governance in concrete deployment settings, especially through case studies of federated clinical AI, procurement templates, hospital safety committees, and post-market reporting workflows. Empirical work should ask whether the proposed triggers are usable, whether logs create actionable evidence rather than compliance noise, whether escalation routes are followed, and whether patients can actually use the redress pathway without technical expertise.

Interface governance is not governance in general. It is the accountability layer required once artifact-centered governance becomes insufficient under conditions of continuous cross-institutional adaptation. Federated AI-SaMD can remain privacy-preserving and clinically adaptive only if it is also contestably reconstructable. The interface, not the model alone, is where that reconstructability must be built.

Ethical, Positionality, and Adverse Impact Statement

The paper does not involve human subjects research or patient data; the arguments are normative and institutional. A central adverse impact concern is that the framework may be appropriated as a formalist compliance checklist: interfaces may be named without being operationalized to support patient redress. The framework mitigates that risk by defining threshold criteria for governance interfaces, specifying triggers, logs, escalation routes, and response duties, and developing a patient-facing contestability pathway. A second concern is that interface obligations may burden smaller hospitals, safety-net providers, and institutions serving underrepresented populations. The framework therefore emphasizes scaled obligations, shared monitoring infrastructure, and sponsor or federation-coordinator support. A third concern is positionality: the argument is developed from a regulatory and AIES-oriented vantage point. Clinicians, patients, community organizations, and lower-resourced institutions would surface practical barriers that a conceptual framework cannot fully capture.

References

- Abbott, A. 1988. *The System of Professions: An Essay on the Division of Expert Labor*. University of Chicago Press.
- Ananny, M.; and Crawford, K. 2018. Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. *New Media & Society*, 20(3): 973–989.
- Bagave, P.; Westberg, M.; Janssen, M.; and Ding, A. Y. 2025. Accountability Framework for Healthcare AI Systems: Towards Joint Accountability in Decision Making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 279–291. DOI: 10.1609/aies.v8i1.36548.
- Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; and Shmatikov, V. 2020. How to Backdoor Federated Learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*.
- Bottomly, D.; Barnes, B.; Mavuwa, K.; Lee, N.; Roth, H. R.; Chen, C.; and McWeeney, S. K. 2026. A Pragmatic Framework for Federated Learning Risk and Governance in Academic Medical Centers. *JMIR AI*, 5: e80022. DOI: 10.2196/80022.
- Bowker, G. C.; and Star, S. L. 1999. *Sorting Things Out: Classification and Its Consequences*. MIT Press.
- Eden, R.; Chukwudi, I.; Bain, C.; Barbieri, S.; Callaway, L.; de Jersey, S.; George, Y.; Gorse, A.-D.; Lawley, M.; Marendy, P.; McPhail, S. M.; Nguyen, A.; Samadbeik, M.; and Sullivan, C. 2025. A Scoping Review of the Governance of Federated Learning in Healthcare. *npj Digital Medicine*, 8: 427. DOI: 10.1038/s41746-025-01836-3.
- Edwards, L.; and Veale, M. 2017. Slave to the Algorithm? Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review*, 16: 18–84.
- Edwards, P. N. 2010. *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. MIT Press.
- Elish, M. C. 2019. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, 5: 40–60.
- European Union. 2024. Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the European Union.

- FDA. 2024. Marketing Submission Recommendations for a Pre-determined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Guidance for Industry and Food and Drug Administration Staff. Final guidance.
- FDA; Health Canada; and MHRA. 2021. Good Machine Learning Practice for Medical Device Development: Guiding Principles. Joint regulatory guidance.
- Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; Schafer, B.; Valcke, P.; and Vayena, E. 2018. AI4People: An Ethical Framework for a Good AI Society. *Minds and Machines*, 28: 689–707.
- Gailmard, L. A.; Spence, D.; Lawrence, C.; and Ho, D. E. 2025. Known Unknowns and Unknown Unknowns: Designing a Scalable Adverse Event Reporting System for AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 1004–1017. DOI: 10.1609/aies.v8i2.36607.
- Galison, P. 1997. *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press.
- Ganesh, B. 2023. The Role of Governance in Bridging AI Responsibility Gaps: An Interdisciplinary Evaluation of Emerging AI Governance Measures. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 952–954. DOI: 10.1145/3600211.3604751.
- Ganesh, B.; Schiff, D. S.; and Anderson, S. 2025. The ‘Wild West’ of Medicine: Exploring the Emergence of ‘Grassroots’ AI Governance in Radiology. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 1018–1031. DOI: 10.1609/aies.v8i2.36608.
- International Medical Device Regulators Forum. 2017. Software as a Medical Device (SaMD): Clinical Evaluation. IMDRF/SaMD WG/N41 FINAL:2017.
- Jobin, A.; Ienca, M.; and Vayena, E. 2019. The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1: 389–399.
- Kaminski, M. E. 2019. Binary Governance: Lessons from the GDPR’s Approach to Algorithmic Accountability. *Southern California Law Review*, 92: 1529–1616.
- Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, 37(3): 50–60.
- March, J. G.; and Olsen, J. P. 1989. *Rediscovering Institutions: The Organizational Basis of Politics*. Free Press.
- Matthias, A. 2004. The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, 6(3): 175–183.
- Mittelstadt, B. 2019. Principles Alone Cannot Guarantee Ethical AI. *Nature Machine Intelligence*, 1: 501–507.
- Nissenbaum, H. 1996. Accountability in a Computerized Society. *Science and Engineering Ethics*, 2: 25–42.
- Rieke, N.; Hancox, J.; Li, W.; Milletari, F.; Roth, H. R.; Albarqouni, S.; Bakas, S.; Galtier, M. N.; Landman, B. A.; Maier-Hein, K.; Ourselin, S.; Sheller, M.; Summers, R. M.; Trask, A.; Xu, D.; Baust, M.; and Cardoso, M. J. 2020. The Future of Digital Health with Federated Learning. *npj Digital Medicine*, 3: 119. DOI: 10.1038/s41746-020-00323-1.
- Santoni de Sio, F.; and Mecacci, G. 2021. Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them. *Philosophy & Technology*, 34: 1057–1084.
- Schiff, D.; Borenstein, J.; Biddle, J.; and Laas, K. 2021. AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection. *IEEE Transactions on Technology and Society*, 2(1): 31–42.
- Scott, W. R. 2014. *Institutions and Organizations: Ideas, Interests, and Identities*. SAGE Publications, 4 edition.
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.
- Sheller, M. J.; Edwards, B.; Reina, G. A.; Martin, J.; and Bakas, S. 2020. Federated Learning in Medicine: Facilitating Multi-Institutional Collaborations without Sharing Patient Data. *Scientific Reports*, 10: 12598. DOI: 10.1038/s41598-020-69250-1.
- Star, S. L.; and Ruhleder, K. 1996. Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research*, 7(1): 111–134.
- Suchman, L. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge University Press.
- Zhu, L.; Liu, Z.; and Han, S. 2019. Deep Leakage from Gradients. In *Advances in Neural Information Processing Systems*.