

ORIGINAL RESEARCH

Open access

Explaining Without Exposing: A Federated Framework for Privacy-Preserving Learning, Explanation Sharing, Confidentiality Protection, and Cross-Organizational Accountability

Laura Mendez^{1*}, Hiroshi Sato², Anna Kowalczyk³

Abstract

Federated learning enables multiple participants to develop shared models without routinely transferring their raw data to a central repository, yet data localization alone does not resolve the trustworthiness challenges of distributed artificial intelligence. Explainability can support understanding, oversight, auditability, and accountability, but explanations may also expose sensitive attributes, local data characteristics, model information, organizational knowledge, provenance traces, or security-relevant evidence. This article proposes a conceptual federated explainable AI framework for integrating privacy-preserving learning, explanation sharing, confidentiality protection, leakage-risk assessment, and cross-organizational accountability. The framework begins by defining the federation context, participants, responsibilities, data boundaries, and sovereignty conditions. It then structures privacy–explainability tension assessment, local and global explanation requirements, explanation granularity and abstraction, access control, redaction, filtering, provenance, model-update review, threat modelling, conditional privacy safeguards, auditability, incident response, and governance updating. The analysis distinguishes local, global, personalized, organization-specific, and cross-organizational explanations while emphasizing that each may produce different disclosure and accountability risks. It also identifies confidentiality threats arising from model updates, gradients, feature attributions, explanation aggregation, membership inference, reconstruction, and proprietary-information leakage. The central contribution is an explanation-boundary approach that asks what may be explained, to whom, for what purpose, at what level of detail, under which safeguards, and with what residual-risk communication. The framework may support responsible design and evaluation, but it does not independently establish privacy, security, confidentiality, legal compliance, ethical sufficiency, accountability, certification, or deployment readiness.

Keywords Explainable AI, Federated learning, Privacy-preserving AI, Explanation sharing, Confidentiality, Leakage risk

*Correspondence:

Laura Mendez
laura.mendez@ie.edu

¹ Department of Human-Centered AI and Explainability, IE Business School, Madrid, Spain

² Department of AI and Decision Science, Keio University, Tokyo, Japan

³ Department of Computer Science and AI, University of Warsaw, Warsaw, Poland

Introduction

Explainable artificial intelligence encompasses methods and practices intended to make model behaviour, predictions, internal relationships, or decision logic more understandable to relevant stakeholders. Its responsible use requires more

than producing an attribution map, rule, example, or natural-language rationale, because explanatory value depends on fidelity, intelligibility, purpose, audience, and the consequences of disclosure. These requirements become especially difficult in federated environments, where knowledge is distributed across organizations and explanatory evidence may cross technical and institutional boundaries. Explainability must therefore be considered alongside privacy, confidentiality, security, and accountability rather than as an isolated transparency function [1].

Explanations are also social and contextual artefacts. A useful explanation for a model developer may be inappropriate for an affected person, local professional, consortium participant, auditor, or security reviewer. Stakeholders may seek different contrasts, levels of detail, causal accounts, examples, or evidence about uncertainty and limitations. In a federation, these differences are intensified because recipients may hold unequal authority, information, expertise, and access rights. Explanation design must consequently specify not only what a model can reveal, but also what each recipient needs to know and what information should remain protected [2].

Federated learning distributes model training across data-holding participants and commonly coordinates local computation through exchanges of parameters, gradients, model deltas, or related statistics. It differs from centralized learning because raw data may remain under local control, yet the coordinating server, client updates, aggregation process, global model, and released outputs still create information flows. Federated AI should therefore not be described as private merely because datasets are not routinely pooled. Its trustworthiness depends on the federation architecture, participant assumptions, communication design, threat model, safeguards, and governance arrangements [3].

The objective of this article is to propose a federated explainable AI framework that integrates privacy-aware explanation sharing, confidentiality protection, evidence traceability, leakage-risk review, and cross-organizational accountability. The framework treats ethical principles as orientation points rather than proof that responsible practices have been achieved. It organizes the questions that participating organizations should examine while avoiding prescriptive claims about particular algorithms, privacy mechanisms, governance bodies, or implementation choices. Its contribution is a structured explanation-boundary model for determining what may be explained, to whom, under which conditions, and with what documented limitations [4].

Privacy–Explainability Tensions in Distributed AI

The first tension concerns transparency versus confidentiality. Greater visibility into model logic may improve scrutiny, but disclosure can reveal proprietary structures, sensitive decision criteria, security weaknesses, or information about the data on which the model depends. Post-hoc explanations are particularly vulnerable to overinterpretation when they are presented as faithful substitutes for an inherently understandable decision process. In high-consequence settings, the availability of an explanation should not be treated as evidence that the underlying model is sufficiently interpretable, appropriate, or accountable [5].

A second tension concerns explanation scope. Local explanations describe particular predictions or nearby model behaviour, whereas global explanations seek to characterize broader patterns. Personalized and organization-specific explanations may increase relevance, while cross-organizational summaries may support shared oversight. However, increasing detail can disclose sensitive attributes, distinctive feature relationships, rare cases, or institutional distributions. Explanation methods also differ in fidelity, stability, representation, and abstraction. Selecting among them therefore requires a task-specific assessment of explanatory usefulness rather than an assumption that one method or level of granularity is universally preferable [6].

Privacy-preserving explanation requires separate attention from privacy-preserving training. An explanation may expose information even when the learning process uses data localization, aggregation, encryption, or differential privacy. Local insight can reveal local-data characteristics; personalized explanations can increase re-identification risk; and aggregated feature importance can disclose participant-specific distributions. Evidence traceability can strengthen auditability while simultaneously revealing data, client, model, or update provenance. Privacy-aware design must therefore examine the

explanation as an independent information product with its own recipients, attack surface, utility requirements, and disclosure constraints [7].

Explanatory outputs may also facilitate extraction of model behaviour or proprietary knowledge. Repeated queries, detailed feature attributions, examples, rules, confidence information, or counterfactual responses can provide an adversary with information that would not be available from predictions alone. In federated settings, the resulting risk may extend beyond a single model owner to multiple participants whose local knowledge contributes to the shared model. Explanation access control, abstraction, redaction, filtering, and monitoring may reduce exposure, but they cannot establish that explanation sharing is safe under every attacker, recipient, or reuse scenario [8].

Federated Learning and Cross-Organizational Intelligence

Federated learning is a distributed learning arrangement in which participating clients conduct local computation and contribute information to a coordinating process. Federated AI is broader, covering the models, explanations, governance relationships, interfaces, and organizational decisions surrounding that process. Distributed AI may include architectures that are not federated, while cross-organizational AI emphasizes collaboration among legally or operationally distinct parties. Collaborative intelligence describes the shared analytic capability that emerges, but it should not obscure the security and privacy risks created by clients, servers, communications, updates, and aggregation [9].

In cross-silo federation, participants are commonly identifiable organizations with relatively stable roles, whereas cross-device federation may involve larger and more dynamic client populations. Each configuration creates different expectations for authentication, participation, monitoring, explanation access, and responsibility. Cross-organizational learning can support collaboration where direct data pooling is unsuitable, but it also creates dependencies involving technical compatibility, data quality, institutional authority, risk tolerance, and governance. These dependencies require explicit coordination rather than an assumption that a shared model automatically creates a shared understanding of responsibility [10].

The local client, local model, global server, and global model perform distinct functions. Local clients process institutional data and generate model updates; the coordinating layer aggregates permitted contributions and distributes revised model states. Gradients, parameters, deltas, metrics, and diagnostic information may be necessary for learning, yet each can contain information about local observations or model behaviour. Privacy and security controls should therefore be evaluated against the actual information exchanged, the knowledge available to potential adversaries, and the possibility that endpoints or authorized participants may behave unexpectedly [11].

Data localization can reduce the routine movement of raw records, but it is not equivalent to data minimization, confidentiality, or data sovereignty. Minimization asks whether each transmitted update, statistic, explanation, or log entry is necessary for the stated purpose. Sovereignty concerns the authority under which institutional data and derived information are processed, transferred, retained, or reused. Multi-institutional federation demonstrates that collaborative learning may occur without centralized raw-data collection, but it does not establish that model updates, outputs, or explanations are free from disclosure risk [12].

Federated governance must allocate responsibility for client operation, server coordination, model ownership, data stewardship, update approval, explanation release, monitoring, and incident escalation. Consortium governance may additionally be needed where participants jointly define rules, admit or remove members, resolve disputes, or change the federation. These arrangements must reject centralized assumptions that a single actor has complete visibility or control. They must also preserve a clear boundary: federated architecture may support data localization and collaborative intelligence, but it does not guarantee privacy. **Table 1** summarises privacy–explainability tensions and federated learning requirements that shape cross-organizational explanation design.

Table 1. Privacy–Explainability Tensions and Federated Learning Requirements in Cross-Organizational AI: Local Clients, Global Models, Client Updates, Data Localization, Data Minimization, Data Sovereignty, Transparency, Confidentiality, Explanation Detail, Evidence Traceability, Auditability, Proprietary Protection, and Federated-Privacy Boundaries

Tension or federated requirement	Core governance question	Explainability or learning function	Privacy or confidentiality risk	Potential safeguard	Failure or overclaiming risk	Governance or evaluation need	Framework
Transparency versus confidentiality	What must be disclosed for legitimate scrutiny?	Supports understanding and review.	Model logic or sensitive knowledge may be exposed.	Tiered disclosure and abstraction.	Treating maximum transparency as inherently trustworthy.	Recipient- and purpose-specific review.	Defining disclosure boundaries
Local insight versus local-data exposure	Can local behaviour be explained without revealing local records or distributions?	Supports local validation.	Rare attributes or sensitive correlations may be inferred.	Redaction, aggregation, and minimum-detail rules.	Assuming locally generated explanations are private.	Local leakage testing.	Local explanation generation
Global explanation versus participant anonymity	Can shared behaviour be described without identifying influential clients?	Supports system-level understanding.	Client contribution or composition may be inferred.	Coarse aggregation and contribution masking.	Equating aggregation with anonymity.	Client-level inference assessment.	Global release
Explanation detail versus leakage risk	What is the minimum sufficient level of detail?	Improves diagnosis and contestability.	Fine-grained evidence increases disclosure.	Granularity and abstraction controls.	Assuming more detail is always better.	Joint usefulness and privacy evaluation.	Explanation boundaries and details
Traceability versus confidentiality	Which provenance elements are necessary for review?	Supports reconstruction of model and explanation history.	Logs may identify protected data, clients, or updates.	Segregated, role-restricted provenance.	Treating complete traceability as harmless.	Retention and access review.	Evidence architecture
Auditability versus proprietary protection	What evidence can an auditor access without acquiring trade secrets?	Enables independent scrutiny.	Proprietary models or processes may be exposed.	Scoped evidence packages and protected review.	Using confidentiality to block all scrutiny.	Auditor-role and necessity assessment.	Audit-boundaries

Local clients and local models	Who controls local training, data, and explanations?	Perform institution-specific learning.	Compromised clients may leak or manipulate information.	Client authentication and local assurance.	Assuming all participants are equally trustworthy.	Client monitoring and incident reporting.	Local responsibility
Global server and global model	Who controls aggregation and global releases?	Coordinates learning across participants.	Server access may create concentration and inference risks.	Separation of duties and controlled aggregation.	Treating the server as a neutral component.	Server assurance and oversight.	Global responsibility
Client, model, gradient, and parameter updates	Which learning signals are necessary to transmit?	Enable global model development.	Reconstruction, inference, poisoning, or proprietary leakage.	Minimization and mechanism-specific safeguards.	Describing updates as non-sensitive metadata.	Update-level threat evaluation.	Update review stage
Data localization and minimization	What remains local, and what derived information leaves the client?	Limits centralized raw-data collection.	Excessive updates, telemetry, or explanations remain possible.	Data-flow inventory and purpose limitation.	Claiming localization proves privacy.	Periodic information-flow audit.	Federated enforcement
Data sovereignty	Which organization controls processing, reuse, and disclosure?	Preserves institutional authority over information.	Derived information may escape original restrictions.	Provenance and transfer controls.	Considering sovereignty only at raw-data level.	Organization-specific governance review.	Participatory control
No federated-privacy guarantee boundary	What residual risks remain after federation and safeguards?	Constrains privacy claims.	Stakeholders may assume decentralized training is sufficient.	Explicit residual-risk communication.	"Data remain local, therefore the system is private."	Lifecycle reassessment.	Management and guarantee boundary

Local and Global Explanation Sharing

Local explanation concerns a particular client, model, case, or decision context, whereas global explanation seeks to characterize shared behaviour across the federation. Federated explainable AI must coordinate these forms without collapsing them into a single explanatory product. A local professional may need institution-specific evidence, while consortium participants may require a limited account of global behaviour. Affected stakeholders, auditors, model owners, and security reviewers may each need different views. Federated XAI therefore requires purposeful explanation design across technical and organizational boundaries [13].

Explanation generation can occur through interpretable local models, post-hoc techniques, aggregated explanatory structures, or global summaries derived from shared learning. The resulting artefacts are not equivalent to gradients, parameters, predictions, or raw data, but they may encode information from each. Explanation aggregation must

document how local evidence is normalized, weighted, filtered, and combined, particularly where participant data are heterogeneous. A global explanation should not be interpreted as representing every local context, and a local explanation should not be generalized beyond the conditions under which it was produced [14].

Personalized explanations may improve relevance for particular recipients, while organization-specific explanations may help institutions examine local errors, subgroup behaviour, or operational suitability. However, personalization can combine attributes in ways that increase re-identification or sensitive-inference risk. Organization-specific outputs can reveal case mix, feature prevalence, workflow characteristics, or proprietary knowledge. Domain applications of federated interpretable models show that local and global explanatory structures can be generated, but their usefulness, fidelity, confidentiality, and transferability require separate evaluation in each setting [15].

Cross-organizational explanation sharing also raises questions of consistency. Comparable explanations may help participants examine whether a global model behaves coherently across clients, but consistency does not prove that an explanation is faithful, useful, private, or safe to release. A federated post-hoc method may coordinate explanation generation while leaving unresolved questions about repeated queries, client contribution, feature-distribution inference, and recipient reuse. Explanation-sharing mechanisms should therefore be evaluated as information-release processes rather than presumed to inherit the privacy properties claimed for the learning architecture [16].

A defensible explanation-sharing process should define the recipient, purpose, permitted content, granularity, abstraction, access rights, retention conditions, and onward-use restrictions before release. Role-based and need-to-know access can limit exposure, while redaction and filtering can suppress identifiers, rare features, sensitive examples, or unnecessary provenance. Explanation provenance and traceability should link an artefact to its method, model version, update cycle, and approval without automatically revealing protected clients or data. These controls may support accountability, but neither restricted sharing nor a complete evidence trail is sufficient by itself to establish confidentiality, privacy, or responsible use [7].

Confidentiality, Leakage, and Security Risks

Differential privacy may constrain the influence of individual training records under formally defined assumptions, but its usefulness in federated learning depends on the protected unit, mechanism, privacy accounting, release composition, and resulting utility loss. A privacy budget is meaningful only within a specified design and cannot be transferred uncritically from model training to explanation release. Explanations, provenance records, and audit outputs may remain outside the mechanism's protection. Differential privacy should therefore be treated as a conditional safeguard requiring configuration-specific validation, not as proof that a federated model or its explanations are free from disclosure risk [17].

Model updates and gradients may reveal information about local training data. Gradient-inversion and reconstruction attacks illustrate that transmitted learning signals can, under particular assumptions, expose approximate inputs, labels, attributes, or other sensitive characteristics. Practical severity varies with model architecture, batch size, attacker knowledge, update structure, and available auxiliary information. Nevertheless, the possibility of reconstruction means that gradients, parameters, and model deltas should not be characterized as harmless technical metadata merely because raw records remain within local systems [18].

Membership inference presents a related risk by seeking to determine whether a record or participant contributed to model training. Such disclosure may itself be sensitive, especially where participation implies a health condition, institutional relationship, behavioral category, or protected activity. Explanations may increase the available attack information by revealing feature influence, confidence, local behavior, or comparative responses. Defense evaluation should therefore consider both ordinary model access and explanation-augmented access, including repeated queries and unequal exposure among rare or influential records [19].

Gradient leakage is not one uniform event but a family of threats involving different adversaries, assumptions, targets, and reconstruction techniques. In federated explainable AI, the attack surface may extend beyond updates to feature attributions, local and global explanations, aggregation outputs, model and client provenance, evidence trails, and audit logs. These artefacts may expose feature distributions, proprietary rules, organizational characteristics, or security weaknesses even where raw-data reconstruction is unsuccessful. **Figure 1** illustrates the privacy–explainability risk pathway in federated AI, showing how local data, model updates, explanations, aggregation, access control, leakage risks, and accountability mechanisms interact [20].

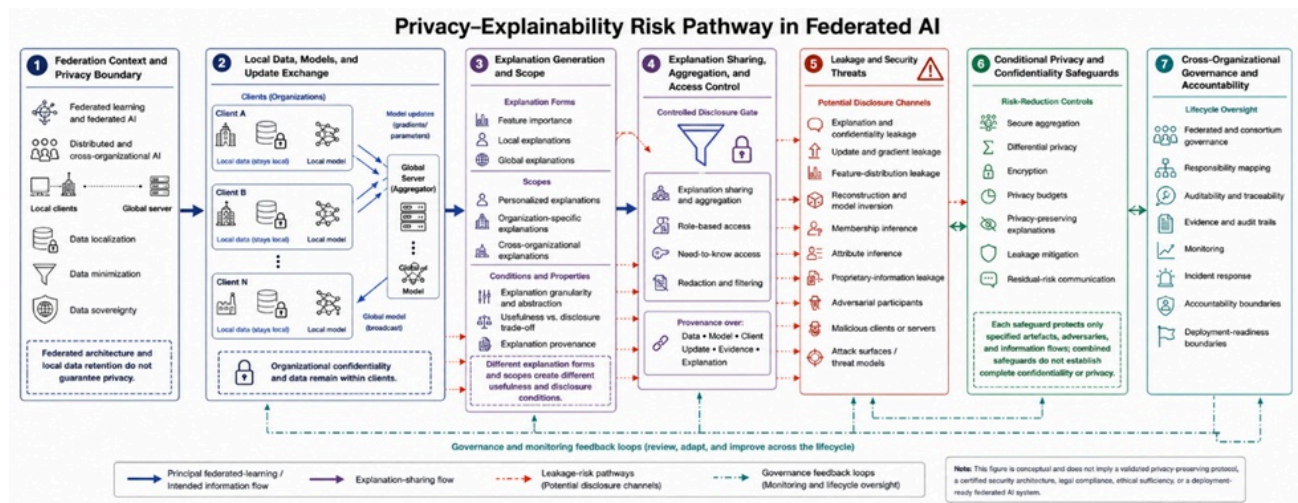


Figure 1. Privacy–Explainability Risk Pathway in Federated AI

Alt-text description

A conceptual pathway showing local clients, local data, model updates, explanation sharing, aggregation, access control, leakage risks, confidentiality safeguards, auditability, and cross-organizational accountability.

Secure aggregation, encryption, and differential privacy may protect particular communications or computations, but none independently covers every endpoint, explanation, log, recipient, metadata flow, or collusion scenario. Evidence that one mechanism reduces gradient exposure does not establish that the same configuration protects against every reconstruction or explanation-assisted attack. Safeguard review must identify precisely which artefacts and adversaries fall within the protection scope, document uncovered risks, and monitor changing attack conditions. Residual risk remains even after multiple controls are combined [21].

Governance and Accountability Requirements

Federated explainable AI requires human-centered governance because technical coordination does not determine who should exercise judgment, accept responsibility, or intervene when explanations are misleading or unsafe. Local professionals, affected stakeholders, model owners, clients, server operators, data stewards, and auditors may require different forms of control and evidence. Human oversight should be assigned to identifiable functions with sufficient authority, information, competence, and escalation routes rather than invoked as a general principle without operational ownership [22].

Cross-organizational accountability requires responsibility mapping across the federation lifecycle. Client operators may be responsible for local data, training, validation, and security; server operators for aggregation and global-model management; model owners for purpose, versions, limitations, and retirement; and explanation providers for method selection, content, and release conditions. Explanation recipients also bear responsibilities concerning authorized use,

retention, interpretation, and onward sharing. Without explicit allocation, distributed participation can produce accountability gaps in which every organization attributes failures to another actor [23].

Transparency alone cannot close those gaps. An organization may disclose model documentation, explanations, or logs without enabling meaningful understanding, challenge, correction, or remedy. Conversely, unlimited visibility may compromise personal information, institutional confidentiality, security, or proprietary knowledge. Auditability should therefore be purpose-specific and role-sensitive, providing authorized reviewers with sufficient evidence while restricting unnecessary exposure. The existence of an explanation, evidence trail, or transparent interface should not be represented as proof that accountability has been established [24].

Documentation supports traceability by recording data provenance, model versions, update cycles, explanation methods, assumptions, limitations, approvals, and changes. Data documentation can clarify collection context, composition, intended use, access constraints, and known limitations, while model and decision documentation can connect technical artefacts to organizational judgments. These records require governance because incomplete, stale, inaccurate, or overexposing documentation can create a false appearance of assurance. Provenance must be usable for investigation without automatically disclosing protected data or participant identities [25].

Federated and consortium governance should define participation, decision rights, explanation access, change control, monitoring, dispute resolution, withdrawal, incident response, and risk communication. Stakeholder participation may help identify explanation needs and unacceptable disclosure, but it should be substantive and appropriately protected. Governance tools may support translation of broad principles into reviewable practices, yet their presence does not establish legal compliance, ethical sufficiency, or effective accountability. Table 2 maps local and global explanation sharing, confidentiality protection, leakage risks, security threats, access control, governance, and accountability requirements for federated explainable AI [26].

Table 2. Explanation Sharing, Confidentiality Protection, Leakage Risk, Security, Governance, and Accountability Requirements for Federated Explainable AI: Local Explanations, Global Explanations, Explanation Aggregation, Granularity, Access Control, Redaction, Provenance, Model Update Leakage, Gradient Leakage, Inference Attacks, Auditability, Traceability, and Accountability Boundaries

Explanation, risk, or governance requirement	Core explanation or security question	Required evidence or process	Privacy, explanation, or accountability function	Potential output	Failure or residual risk	Access, monitoring, or governance need	Framework
Local explanation	What local evidence is necessary?	Fidelity, usefulness, and leakage review.	Support local understanding.	Restricted local rationale.	Local-data or attribute exposure.	Local role-based access.	Local explanation channels
Global explanation	What shared behavior can be disclosed?	Aggregation and heterogeneity analysis.	Support consortium understanding.	Global explanatory summary.	Client inference or false generalization.	Consortium approval.	Global explanation channels
Personalized or organization-specific explanation	What tailoring is justified?	Purpose and re-identification review.	Improve recipient relevance.	Tailored explanation.	Sensitive or institutional disclosure.	Authenticated isolated access.	Consortium explanation channels

Explanation aggregation	How are local explanations combined?	Weighting, normalization, and contribution review.	Construct shared explanatory evidence.	Aggregated rules or attributions.	Dominant-client bias or contribution leakage.	Aggregation audit.	Explanation consistency
Granularity and abstraction	What is the minimum sufficient detail?	Joint user and privacy testing.	Balance usefulness and exposure.	Tiered explanation views.	Overexposure or unusable abstraction.	Recipient-specific policy.	Boundaries of de-mechanization
Access control	Who may request, receive, or export explanations?	Identity, role, purpose, and authorization.	Enforce least privilege.	Access decision and log.	Privilege escalation or onward sharing.	Entitlement monitoring.	Disclosure controls
Redaction and filtering	Which information should be suppressed?	Disclosure and materiality analysis.	Reduce sensitive content.	Filtered explanation package.	Incomplete redaction or misleading omission.	Pre-release review.	Explanation safety
Provenance and traceability	Can the explanation be linked to its origin and transformations?	Versioned protected metadata.	Support audit and correction.	Explanation evidence chain.	Provenance itself exposes clients or data.	Segregated access.	Evidence of provenance
Update and gradient leakage	Can learning signals reveal local information?	Threat-specific testing.	Protect training information.	Update-risk assessment.	Reconstruction or proprietary leakage.	Security monitoring.	Update review
Membership, inversion, and attribute inference	Can participation or sensitive properties be inferred?	Model- and explanation-aware attack testing.	Protect records and participants.	Inference-risk statement.	Unequal protection and auxiliary-data attacks.	Repeated-query monitoring.	Privacy leakage
Secure aggregation, differential privacy, and encryption	What does each mechanism actually protect?	Scope, configuration, and residual-risk review.	Reduce specified exposures.	Safeguard-suitability record.	Endpoint, metadata, explanation, or collusion risk.	Specialist validation.	Confidentiality and safety
Auditability and evidence trails	What evidence is needed for legitimate scrutiny?	Scoped audit plan and controlled records.	Enable independent review.	Audit evidence package.	Audit theatre or excessive disclosure.	Auditor-role governance.	Accountability and evidence
Responsibility mapping	Who owns each lifecycle decision and failure?	Actor, duty, authority, and escalation map.	Prevent responsibility diffusion.	Responsibility register.	Unowned controls.	Periodic governance review.	Accountability foundation

Accountability and compliance boundaries	What remains unproven?	Explicit residual-gap statement.	Constrain claims.	Qualified governance statement.	Procedures represented as certification.	Claim review.	Man n gua bou
--	------------------------	----------------------------------	-------------------	---------------------------------	--	---------------	------------------------

Proposed Federated Explainable AI Framework

The proposed framework begins with use-case and federation-context definition. It records the decision purpose, federation architecture, participant type, local and global models, information flows, data distribution, coordination process, and intended recipients of model and explanation outputs. This stage distinguishes cross-silo, cross-device, bilateral, and consortium arrangements where relevant. It also rejects the centralized assumption that one actor necessarily possesses complete evidence or authority. System-level research shows that architecture, scale, participation, and data distribution materially shape privacy and governance requirements [27].

The second stage maps participants and responsibilities, followed by data-localization and sovereignty review. Clients, server operators, model owners, data stewards, explanation providers, explanation recipients, auditors, and affected stakeholders are assigned reviewable duties. The framework then identifies protected data, institutional information, proprietary knowledge, update signals, explanatory artefacts, and provenance records. This multidimensional approach reflects trustworthy federated learning as a combination of privacy, security, robustness, explainability, accountability, and governance rather than a privacy-only technical arrangement [28].

The third stage assesses privacy–explainability tensions and defines local and global explanation requirements. It asks which stakeholder needs an explanation, for what task, at what scope and abstraction, and with which limitations. Personalized, organization-specific, and cross-organizational explanations are included only where supported by purpose and risk evidence. Domain-adaptive explanation is essential because an explanation useful to a model developer may not meet the informational, cognitive, professional, or accountability needs of another recipient [29].

The fourth stage designs explanation granularity, access control, redaction, filtering, and provenance, then reviews model updates, leakage risks, and the threat model. Secure aggregation, differential privacy, and encryption may be considered where supported, but each is evaluated against its actual protection scope. Cross-organizational auditability is designed through controlled evidence access rather than unrestricted disclosure. Technical architectures may help preserve records and attribute actions, but architectural traceability does not independently establish fairness, accountability, or legitimate governance [30].

The final stage establishes monitoring, incident response, risk communication, governance review, and framework updating. New participants, model changes, explanation methods, attack findings, audit results, or stakeholder concerns may require the framework to be revisited. Privacy, security, and trustworthiness remain lifecycle properties subject to changing contexts and residual risks. **Table 3** presents the proposed federated explainable AI framework linking federation context, privacy–explainability assessment, local and global explanations, access control, leakage review, provenance, auditability, accountability, governance review, and framework updating [31].

Table 3. Proposed Federated Explainable AI Framework for Privacy-Preserving Learning and Cross-Organizational Accountability: Federation Context, Participant Mapping, Data Localization, Privacy–Explainability Tensions, Local Explanation Requirements, Global Explanation Requirements, Explanation Granularity, Access Control, Redaction, Provenance, Leakage Risk Assessment, Auditability, Governance Review, and Deployment Boundaries

Framework stage	Core purpose	Required input	Privacy-explainability function	Potential output	Validation or governance need	Failure risk
-----------------	--------------	----------------	---------------------------------	------------------	-------------------------------	--------------

Federation-context definition	Define purpose, architecture, and participants.	Use case and system map.	Prevent generic assumptions.	Federation-context record.	Multi-party review.	Ill-defined scope.
Participant and responsibility mapping	Assign duties and authority.	Participant and role registry.	Link access to responsibility.	Responsibility map.	Governance approval.	Responsibility diffusion.
Data-localization and sovereignty review	Identify control over data and derivatives.	Data flows, authority, location, provenance.	Separate localization from privacy proof.	Sovereignty conditions.	Data-steward review.	Unauthorized transfer or reuse.
Privacy–explainability assessment	Identify competing benefits and risks.	Stakeholder needs and threat model.	Make trade-offs explicit.	Tension register.	Privacy and explanation review.	One objective dominates unjustifiably.
Local explanation requirements	Define local purpose and recipient.	Local workflow and decision needs.	Limit unnecessary local disclosure.	Local explanation specification.	User and leakage evaluation.	Local-data exposure.
Global explanation requirements	Define consortium-wide explanatory needs.	Global model and stakeholder map.	Distinguish shared understanding from participant disclosure.	Global explanation specification.	Consortium approval.	False generalization or client inference.
Granularity and abstraction design	Select minimum sufficient detail.	Utility and sensitivity evidence.	Balance understanding and leakage.	Tiered explanation levels.	Comparative testing.	Excess detail or unusable abstraction.
Access-control design	Restrict request, receipt, export, and reuse.	Role, identity, purpose, and sensitivity.	Enforce least privilege.	Access policy and logs.	Excessive access or onward disclosure.	Entitlement review.
Redaction and filtering	Remove unnecessary sensitive information.	Explanation schema and disclosure analysis.	Reduce identity, feature, and proprietary leakage.	Filtered explanation.	Dual usefulness/privacy review.	Failed redaction or material omission.

Provenance review	Link explanation to model, method, update, and approval.	Versioned protected metadata.	Support audit with controlled disclosure.	Evidence chain.	Audit and confidentiality review.	Missing or overexposing provenance.
Update and leakage review	Examine gradients, parameters, telemetry, explanations, and logs.	Information-flow inventory.	Identify multi-channel disclosure.	Leakage-risk assessment.	Security and privacy evaluation.	Reconstruction or inference.
Threat-model and safeguard review	Match controls to adversaries and artefacts.	Attack surface and safeguard configuration.	Evaluate secure aggregation, DP, and encryption where supported.	Safeguard-scope statement.	Specialist review.	Mechanism overgeneralization.
Auditability and incident planning	Enable scrutiny, containment, and communication.	Audit scope, escalation, and evidence needs.	Balance accountability and confidentiality.	Audit and incident plans.	Multi-party testing.	Audit theatre or delayed response.
Governance review and updating	Maintain contextual validity.	Monitoring, incidents, audits, and stakeholder feedback.	Reassess explanation and privacy boundaries.	Revised framework decision.	Periodic consortium review.	Stale governance.

Practical Implications and Future Research

Privacy-aware explanation design should evaluate explanations as disclosure channels in their own right. Future studies should compare local and global explanation fidelity, stability, usefulness, and leakage across heterogeneous clients rather than reporting only global-model accuracy. Local-sharing research should examine whether explanations reveal rare cases or institutional characteristics, while global-aggregation research should evaluate weighting, normalization, participant influence, and false consensus. Explanation granularity should be tested jointly for human usefulness and privacy risk rather than selected solely for interpretive richness [7].

Access-control and redaction research should address authorized-user misuse, repeated-query composition, onward sharing, and inference from combinations of apparently harmless outputs. Explanation-leakage evaluation should extend beyond model extraction to feature-distribution inference, reconstruction, membership inference, attribute inference, and proprietary-information exposure. Particular attention is needed where explanations are combined with gradients, parameters, confidence values, provenance, or auxiliary information, because multi-channel attacks may reveal more than any single artefact [20].

Research on differential privacy for explanation sharing should specify the protected unit, neighboring relation, privacy accounting, composition, utility effects, and information outside the guarantee. Secure aggregation research should

similarly distinguish protection of individual updates from exposure through endpoints, aggregate outputs, explanations, metadata, and logs. Neither mechanism should be recommended independently of the federation architecture, attacker assumptions, explanation method, and domain-specific risk assessment [21].

Cross-organizational accountability requires empirical study of responsibility maps, audit access, incident exercises, stakeholder participation, evidence quality, and dispute processes. Auditability research should examine how much provenance an authorized reviewer needs and how confidentiality can be preserved without weakening scrutiny. Comparative work is also needed across cross-silo, cross-device, bilateral, and consortium governance arrangements. These are conceptual implications and research priorities, not legal, compliance, procurement, audit, cybersecurity, privacy-engineering, model-selection, platform-selection, vendor-selection, certification, or deployment instructions [26].

Conclusion

Explaining without exposing requires federated AI to treat privacy and explainability as interdependent design and governance problems. The proposed framework links federation context, participant responsibilities, data localization and sovereignty, local and global explanation boundaries, granularity, access control, redaction, provenance, model-update review, leakage-risk assessment, conditional safeguards, auditability, incident response, and governance updating. Its central premise is that neither federated learning nor explanation availability independently establishes responsible AI. Trustworthy cross-organizational intelligence instead requires evidence that explanations are useful for their intended recipients while remaining proportionate to confidentiality, security, proprietary-information, and privacy risks. Residual uncertainty must be documented and communicated, and framework decisions must be revisited as models, participants, threats, and stakeholder needs change. The framework supports structured inquiry and future evaluation, but it does not certify privacy, confidentiality, security, accountability, compliance, ethical sufficiency, or deployment readiness.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 31 Aug 2025

Revised: 26 Sep 2025

Accepted: 10 Nov 2025

Published online: 15 March 2026

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.
<https://doi.org/10.1016/j.inffus.2019.12.012>.
- Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell*. 2019;267:1-38.
<https://doi.org/10.1016/j.artint.2018.07.007>.
- Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. *ACM Trans Intell Syst Technol*. 2019;10(2):1-19.
<https://doi.org/10.1145/3298981>.
- Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell*. 2019;1(9):389-99.
<https://doi.org/10.1038/s42256-019-0088-2>.
- Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-15.
<https://doi.org/10.1038/s42256-019-0048-x>.
- Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv*. 2018;51(5):1-42.
<https://doi.org/10.1145/3236009>.
- Nguyen TT, Huynh TT, Ren Z, Nguyen TT, Nguyen PL, Yin H, et al. Privacy-preserving explainable AI: a survey. *Sci China Inf Sci*. 2025;68(1):111101.
<https://doi.org/10.1007/s11432-024-4123-4>.
- Yan A, Huang T, Ke L, Liu X, Chen Q, Dong C. Explanation leaks: explanation-guided model extraction attacks. *Inf Sci*. 2023;632:269-84.
<https://doi.org/10.1016/j.ins.2023.03.020>.
- Mothukuri V, Parizi RM, Pouriyeh S, Huang Y, Dehghantanha A, Srivastava G. A survey on security and privacy of federated learning. *Future Gener Comput Syst*. 2021;115:619-40.
<https://doi.org/10.1016/j.future.2020.10.007>.
- Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med*. 2020;3(1):119.
<https://doi.org/10.1038/s41746-020-00323-1>.
- Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell*. 2020;2(6):305-11.
<https://doi.org/10.1038/s42256-020-0186-1>.
- Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep*. 2020;10(1):12598.
<https://doi.org/10.1038/s41598-020-69250-1>.

Kusiak A. Federated explainable artificial intelligence (fXAI): a digital manufacturing perspective. *Int J Prod Res.* 2024;62(1-2):171-82.

<https://doi.org/10.1080/00207543.2023.2238083>.

Corcuera Bárcena JL, Ducange P, Marcelloni F, Nardini G, Noferi A, Renda A, et al. Enabling federated learning of explainable AI models within beyond-5G/6G networks. *Comput Commun.* 2023;210:356-75.

<https://doi.org/10.1016/j.comcom.2023.07.039>.

Ducange P, Marcelloni F, Renda A, Ruffini F. Federated learning of XAI models in healthcare: a case study on Parkinson's disease. *Cogn Comput.* 2024;16(6):3051-76.

<https://doi.org/10.1007/s12559-024-10332-x>.

Ducange P, Marcelloni F, Miglionico GC, Renda A, Ruffini F. Federated SHAP: privacy-preserving and consistent post-hoc explainability in federated learning. *Mach Learn.* 2026;115(2):24.

<https://doi.org/10.1007/s10994-025-06956-1>.

Wei K, Li J, Ding M, Ma C, Yang HH, Farokhi F, et al. Federated learning with differential privacy: algorithms and performance analysis. *IEEE Trans Inf Forensics Secur.* 2020;15:3454-69.

<https://doi.org/10.1109/TIFS.2020.2988575>.

Hatamizadeh A, Yin H, Molchanov P, Myronenko A, Li W, Dogra P, et al. Do gradient inversion attacks make federated learning unsafe? *IEEE Trans Med Imaging.* 2023;42(7):2044-56.

<https://doi.org/10.1109/TMI.2023.3239391>.

Bai L, Hu H, Ye Q, Li H, Wang L, Xu J. Membership inference attacks and defenses in federated learning: a survey. *ACM Comput Surv.* 2025;57(4):1-35.

<https://doi.org/10.1145/3704633>.

Gong H, Jiang L, Liu X, Wang Y, Gastro O, Wang L, et al. Gradient leakage attacks in federated learning. *Artif Intell Rev.* 2023;56(Suppl 1):1337-74.

<https://doi.org/10.1007/s10462-023-10550-z>.

Hu J, Du J, Wang Z, Pang X, Zhou Y, Sun P, et al. Does differential privacy really protect federated learning from gradient leakage attacks? *IEEE Trans Mob Comput.* 2024;23(12):12635-49.

<https://doi.org/10.1109/TMC.2024.3417930>.

Shneiderman B. Human-centered artificial intelligence: reliable, safe and trustworthy. *Int J Hum Comput Interact.* 2020;36(6):495-504.

<https://doi.org/10.1080/10447318.2020.1741118>.

Martin K. Ethical implications and accountability of algorithms. *J Bus Ethics.* 2019;160(4):835-50.

<https://doi.org/10.1007/s10551-018-3921-3>.

Ananny M, Crawford K. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc.* 2018;20(3):973-89.

<https://doi.org/10.1177/1461444816676645>.

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H 3rd, et al. Datasheets for datasets. *Commun ACM.* 2021;64(12):86-92.

<https://doi.org/10.1145/3458723>.

Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci Eng Ethics.* 2020;26(4):2141-68.

<https://doi.org/10.1007/s11948-019-00165-5>.

Li Q, Wen Z, Wu Z, Hu S, Wang N, Li Y, et al. A survey on federated learning systems: vision, hype and reality for data privacy and protection. *IEEE Trans Knowl Data Eng.* 2023;35(4):3347-66.
<https://doi.org/10.1109/TKDE.2021.3124599>.

Tariq A, Serhani MA, Sallabi FM, Barka ES, Qayyum T, Khater HM, et al. Trustworthy federated learning: a comprehensive review, architecture, key challenges, and future research prospects. *IEEE Open J Commun Soc.* 2024;5:4920-98.
<https://doi.org/10.1109/OJCOMS.2024.3438264>.

Chaddad A, Lu Q, Li J, Katib Y, Kateb R, Tanougast C, et al. Explainable, domain-adaptive, and federated artificial intelligence in medicine. *IEEE/CAA J Autom Sinica.* 2023;10(4):859-76.
<https://doi.org/10.1109/JAS.2023.123123>.

Lo SK, Liu Y, Lu Q, Wang C, Xu X, Paik HY, et al. Toward trustworthy AI: blockchain-based architecture design for accountability and fairness of federated learning systems. *IEEE Internet Things J.* 2023;10(4):3276-84.
<https://doi.org/10.1109/JIOT.2022.3144450>.

Chen C, Liu J, Tan H, Li X, Wang KIK, Li P, et al. Trustworthy federated learning: privacy, security, and beyond. *Knowl Inf Syst.* 2025;67(3):2321-56.
<https://doi.org/10.1007/s10115-024-02285-2>.