International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 5, Sep-Oct -2026, ISSN (Online): 2395-566X

Responsible Ai Framework for Trustworthy Decision-Making in Cloud Computing

Pawar Pritesh Sandip[1], Mrs. Khushbu Trivedi[2]
Artificial intelligence and data science
Marathwada Mitra mandals institute of technology,Pune

Abstract- This paper presents a design in which every decision produced by a cloud-hosted AI model is examined by a dedicated Responsible AI layer before anyone acts on it. Cloud platforms now run models that classify, predict and recommend in settings where errors are costly, yet an accuracy score alone cannot show whether a particular decision was fair, private, secure or adequately justified. The proposed framework adds an assessment stage downstream of the model. This stage tests each output against fairness, privacy, security and reliability criteria, attaches a feature-level explanation generated with SHAP or LIME, merges the findings into a single risk label, and passes high-risk or uncertain outputs to a human reviewer. Every final decision is written to an audit log, and monitoring follows model behaviour, input drift and responsible-AI risks over time. The paper lays out the architecture, a thirteen-step decision procedure, a compact prototype plan built on Python, Scikit-learn, SHAP, LIME, Fairlearn and a public cloud, and an evaluation plan that contrasts the framework with an unmodified baseline. Because the prototype has not yet been built or tested, no experimental results are reported. The contribution is the integrated design together with a measurable plan for validating it.

Keywords: Responsible AI, Trustworthy AI, Explainable AI, Fairness, Cloud Computing, Decision-Making, Human Oversight, SHAP, LIME

I. INTRODUCTION

AI services running in the cloud now influence outcomes for real people, for example which applicants are approved for credit, which transactions are treated as suspicious and which patients are prioritised for attention. Cloud delivery makes these services easy to scale, yet it also routes data, models and results through shared infrastructure and delivers them to users across a network. If a decision turns out to be wrong, biased or unexplained, the person affected rarely has any means of learning the reason or of appealing it.

In much of current practice a deployed model is judged chiefly by its accuracy. Accuracy is necessary, but it does not show whether results differ between groups, whether personal data was handled safely, whether the model doubted its own answer, or whether the decision can be audited later. Work on explainability, fairness, privacy and security has produced useful techniques for each of these concerns [2], [6], and standardisation efforts such as IEEE P3396[8] consider how AI risk, safety, trustworthiness and responsibility might be defined and evaluated.

In practice, however, these techniques are normally applied in isolation from one another.

1. Contributions

- A Responsible AI assessment layer that evaluates every cloud-generated decision for fairness, privacy, security and reliability.
- An explanation step based on SHAP or LIME that accompanies each decision.
- A risk-based routing rule that sends high-risk or uncertain decisions to a human reviewer and logs every final decision for audit.
- A prototype plan and an evaluation plan that compare the framework with a baseline built on the same model and data.

2. Scope

This work is a design study accompanied by a small planned prototype. It does not train a new large model, does not use confidential data, and does not claim to eliminate every risk of AI decision-making. The framework assembles established techniques into a single pipeline; it is not offered as a new learning algorithm.

II. RELATED WORK

Eight sources dated between 2020 and 2026 were examined: seven research papers and one IEEE recommended practice. As Table 1 shows, each one deals with a different aspect of trustworthy AI.

© 2026 IJSRET

Table 1. Sources reviewed

Ref.	Source	Year	Aspect covered	Role in this paper
[1]	Kaza, explainable agentic-AI NLP for transparent policy decisions in cloud systems	2026	Transparency of policy decisions made in cloud systems	Base paper; motivates transparent cloud decisions
[2]	Wang, Liu, Huang, open API service for trustworthy explanations of cloud AI services	2024	Explanation offered as a cloud service, tested on image classification	Supports the explanation layer
[3]	Trustworthy AI/ML framework for B5G/6G networks	2022	Framework for next-generation networks	Example of a framework-level approach
[4]	AI governance and levels of automation for AIOps-supported administration	2020	Deciding which actions may run automatically and which need people	Informs the routing idea
[5]	Responsible AI, SDGs and AI governance in Africa	2022	Ethics and governance	Governance perspective
[6]	Review of explainable AI and its ethical implications	2024	Building trust and transparency through XAI	Supports using XAI to build trust
[7]	Hybrid-DLT based trustworthy AI framework	2024	Traceability through distributed ledgers	Alternative route to trust
[8]	IEEE P3396 recommended practice	n.d.	Defining and evaluating AI risk, safety, trustworthiness and responsibility	Guides the choice of criteria

.

Read together, these works make three points. First, decisions taken in cloud systems must be explainable and trusted [1], [2], [6]. Second, governance, levels of automation and standards influence how risk and trustworthiness are defined and judged [4], [5], [8]. Third, trustworthy AI is often pursued by means of a framework, whether one aimed at next-generation networks
[3] or one built on distributed ledger technology [7]. Reference [1] serves as the anchor for this study, [2] backs its explanation component, and [8] guides its risk criteria.

III. RESEARCH GAP

Among the sources reviewed, each study concentrates on one part of the problem: explaining cloud AI services, establishing trust and transparency, governing levels of automation, or defining a trustworthy framework for a particular setting. None of them treats the generated decision itself as the object of a complete check inside one cloud workflow, where it is assessed for fairness, privacy, security and reliability, explained, risk-scored, and then sent either to automatic handling or to a person, with the outcome recorded. Four needs follow from this:

- A single integrated framework in place of separate treatments of each responsible-AI concern.
- An assessment carried out after the decision is generated, and applied to that specific decision.
- A stated reason that accompanies every decision.
- Human review for risky or low-confidence decisions, together with an audit record of the final outcome.

1. Background

1. Responsible And Trustworthy Ai

Responsible AI describes the practice of designing and operating AI systems so that they are fair, transparent, explainable, privacy-conscious, secure, reliable and accountable. Trustworthy AI is the result of that practice: a system that people can reasonably depend on. In this paper the seven properties serve as the yardstick against which each cloud decision is judged [6], [8].

2. Explainability With Shap And Lime

SHAP is rooted in the Shapley value concept from cooperative game theory, and it divides a prediction among the input features according to how much each one

© 2026 IJSRET

contributed. LIME takes a different route: it perturbs a single input, observes how the model responds, and trains a small interpretable model on those local results. Both methods work with many model types. Both
also produce approximations, so they help a person understand a decision without proving it correct [2], [6].

2. Fairness And Bias Detection

When training data is biased, outcomes can differ unequally between groups. Fairness-aware learning addresses this by comparing results across groups, for example the difference in positive-decision rates. The Fairlearn library provides such metrics, along with some mitigation methods, in Python. The exact metrics will be chosen once the dataset has been selected.

3. Privacy And Security In The Cloud

Cloud-based AI sends data and models across networks, so both personal information and access to the service need safeguards. Privacy-preserving techniques reduce how much personal data is exposed, while authentication, authorisation and logging help prevent misuse. The framework treats both as assessment criteria: any privacy or security concern found in a decision raises that decision's risk level.

4. Human Oversight And Governance

In a human-in-the-loop arrangement, a person checks or approves selected decisions. This is most useful when the model is uncertain or when a mistake would be costly. Governance adds the requirement that responsibilities are assigned and that decisions can be traced and audited afterwards [4], [5].

5 Deployment And Monitoring

In the cloud, a model runs on managed compute behind a REST API, supported by storage and databases for data and records. Monitoring and logging services follow model performance and shifts in incoming data over time.

6. Conventional Decision-Making Versus The Proposed Framework

Table 2. Conventional approach compared with the proposed framework.

Criterion	Conventional AI decision-making	Proposed framework
Main focus	Prediction accuracy and performance	Accuracy together with fairness, explainability, privacy, security and accountability
Responsible-AI principles	Handled as separate requirements	Built into the cloud decision process
Understanding decisions	Complex models are difficult to verify	SHAP or LIME supplies reasons
Risky decisions	Risk is not always identified	The risk assessment flags high-risk or uncertain cases
Human oversight	No defined review step	High-risk cases are sent to a person
Accountability	Records may be insufficient	Final decisions are logged for audit and monitoring

IV. PROPOSED FRAMEWORK

1. Overview

The framework consists of two linked parts. The first is an ordinary AI decision pipeline: data arrives in the cloud, is prepared, passes through a model, and yields a decision. The second is the Responsible AI assessment layer, which examines that decision for fairness, privacy, security and reliability, explains it, and scores its risk. Depending on the risk, the decision is either handled automatically or passed to a person. The final outcome is recorded, and monitoring results feed back into improvement of the system.

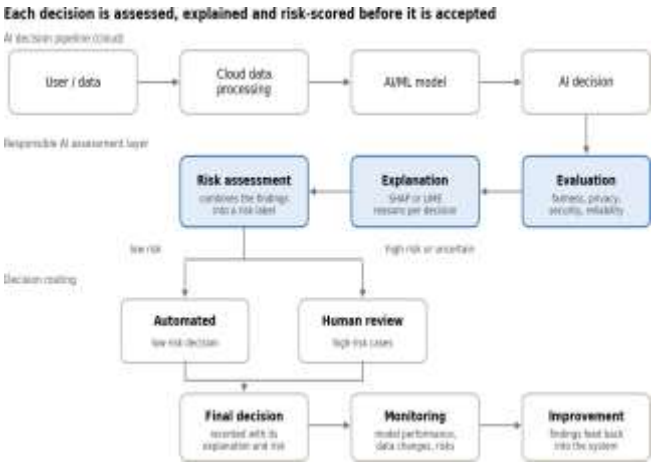


Figure 1. Architecture: white boxes show the AI decision pipeline and its outputs; shaded boxes show the proposed assessment layer.

2. Components

- **User and data:** the origin of the input for which a decision is required.
- **Cloud data processing:** the data is stored and prepared on a cloud platform.
- **Model and decision generation:** a chosen model produces a prediction or decision, along with its confidence where this is available.
- **Responsible AI evaluation:** the decision is checked using fairness metrics, a privacy assessment, security checks and a reliability analysis.
- **Explainability module:** SHAP or LIME identifies the features and factors that lie behind the decision.
- **Risk assessment:** the evaluation results are combined to label the decision as low-risk or high-risk.
- **Routing and human review:** low-risk decisions may be processed automatically, while high-risk or uncertain ones go to a human reviewer.
- **Final decision and audit:** the outcome is stored together with its explanation and risk assessment.
- **Monitoring and improvement:** model performance, data changes and responsible-AI risks are tracked and used to improve the system.

V. METHODOLOGY AND ALGORITHM

1. Research Method

The study proceeds through seven steps.

- **Literature review.** Recent work on responsible AI, trustworthy AI, governance, explainability, fairness, privacy, security and cloud decision-making is gathered from IEEE Xplore, the ACM Digital Library, SpringerLink, ScienceDirect, Google Scholar and arXiv.
- **Study of existing methods.** Explainability, fairness-aware learning, bias detection, privacy-preserving AI, governance and trustworthy-AI frameworks are compared in terms of mechanism, strengths, limits, computational needs and evaluation practice.
- **Cloud decision generation.** A selected model is deployed in a cloud environment using a suitable dataset, and its decisions are examined for bias, uncertainty and reliability.
- **Responsible AI evaluation.** Decisions are assessed individually using fairness metrics, performance measures, a privacy assessment, security checks and a reliability analysis.
- **Explainability layer.** SHAP or LIME identifies the features behind each decision.
- **Risk assessment and human oversight.** The results separate low-risk from high-risk decisions; the latter go to a reviewer, and every final decision is recorded.
- 7. **Evaluation.** The framework is compared with baseline approaches on the parameters listed in Section 8.

2. Risk Assessment And Routing

A decision is marked high-risk if the evaluation reveals strong model uncertainty, unfair outcomes, a privacy issue or any other serious concern. These decisions are sent to a human reviewer together with their explanations, while the remaining ones can be handled automatically. Every outcome is logged either way. The numerical risk score, and the thresholds that define “high-risk” and “low confidence”, have not yet been fixed. They will be set during development on a small development set kept separate from the final test data, so that the thresholds are not tuned on the data used for evaluation. Since the assessment itself relies on metrics and explanation methods with known limits, it lowers risk without removing it.

Algorithm

1. Part A: Set-Up (Performed Once)

- Select the model and a suitable dataset.
- Deploy the model in a cloud environment behind a REST API.
- Define the baseline as the same model without the Responsible AI layer.

2. Part B: For Each Input

- Receive the input and process it in the cloud.
- Have the model generate the decision and, where available, its confidence.
- Evaluate the decision for fairness, privacy, security and reliability.
- Generate an explanation using SHAP or LIME.
- Combine the results of steps 5 to 7 into a risk assessment.
- If the decision is high-risk, go to step 11; otherwise go to step 10.
- Process the low-risk decision automatically and go to step 12.
- Forward the decision and its explanation for human review and adopt the reviewer's final decision.
- Write the final decision, its explanation and its risk assessment to the audit log.
- Monitor performance, data changes and responsible-AI risks, and use the findings for improvement.

© 2026 IJSRET

3. Part C: Evaluation
1. Run the baseline and the framework on the same data and compare them on the parameters in Section 8.

1. Implementation Plan

The prototype is intentionally small and relies only on widely available tools. It requires a laptop with an Intel Core i5 or AMD Ryzen 5 processor or better, 8 GB of RAM (16 GB preferred), 20 GB of free disk space and an internet connection; a GPU is optional. The software stack comprises Windows 10/11 or Ubuntu Linux, Python 3.10 or later, Visual Studio Code, Jupyter Notebook or Google Colab, and Git with GitHub.

Table 3. Prototype tools.

Library or service	Purpose
Pandas, NumPy	Data handling and numerical computation
Scikit-learn	Building the model and computing performance metrics
SHAP, LIME	Explaining individual decisions
Fairlearn	Assessing the fairness of model outcomes
Matplotlib	Plotting results
FastAPI or Flask	Exposing the decision service as a REST API
Requests	Calling the REST APIs from Python
AWS, Azure or Google Cloud (one to be chosen)	Compute, storage, database, authentication, monitoring and logging

The prototype is divided into seven modules: data processing, the decision model, Responsible AI evaluation, the explainability module, risk assessment and routing, audit and monitoring, and an evaluation module that runs the baseline and the framework and records the metrics.
All are planned and none is complete. A small, non-confidential dataset will be used, cloud access keys will be kept out of notebook code, and no real confidential data will be involved.

1. Evaluation Plan
The framework has not been tested, so this section describes how it will be judged and reports no measurements.

1. Design
The baseline and the framework will share the same model and dataset, so any difference can be attributed to the assessment layer. The comparison will also record what the framework costs, such as a higher human-review rate and longer processing time, and will confirm that accuracy remains acceptable. The dataset and the number of test decisions will be fixed when the model is chosen.

2. Parameters

Table 4. Evaluation parameters

Parameter	What is checked
Prediction accuracy	Proportion of correct decisions, to confirm that accuracy stays acceptable
Precision, recall, F1-score	Used where they suit the chosen task
Fairness	Whether outcomes differ across groups, measured with fairness metrics
Explainability	Whether understandable reasons are available for decisions
Reliability	Whether decisions stay consistent under the checks
Privacy	Whether privacy concerns are found in decisions and data handling
Security	Whether the security checks pass for the cloud service
Risk score	The risk assessment produced for each decision
Human-review rate	Proportion of decisions forwarded to a person
Computational cost	Additional resources consumed by the assessment layer
Decision time	Time per decision with and without the assessment layer

3. Expected Behaviour (Hypothetical)
The cases below describe the intended behaviour for a hypothetical approval decision. They are design goals, not observed results.

© 2026 IJSRET

Table 5. Intended handling in three illustrative cases.

Situation	Assessment	Intended handling
The model is confident, and no fairness, privacy or security concern is found	Low risk	Processed automatically and recorded with its explanation
The model is uncertain about the decision	High uncertainty, so high risk	Forwarded for human review before use

Situation	**Assessment**	**Intended handling**
The fairness check shows a possible unfair outcome for a group	Fairness concern, so high risk	Forwarded for human review; the concern is recorded

4. Results

Results are pending. Each parameter in Table 4 will be measured for both the baseline and the framework once the prototype has been built and tested, and the findings will be reported in this section.

5. Applications, Advantages And Limitations

1. **Potential applications**

These are possible uses only; the present work is a small study with a planned prototype.

- **Healthcare:** decisions affecting people can be explained and, when risk is high, reviewed by a person.
- Finance: decisions that are fair, explainable and auditable.
- **Cybersecurity:** AI-driven monitoring whose alerts arrive with clear reasons.
- **Recommendation systems:** transparent and accountable recommendations in cloud applications.
- **Business decision support:** decisions that can be traced, explained and audited, with human oversight where risk is high.

6. Expected Advantages

These follow from the design and remain to be confirmed through testing.

- Fairness, privacy, security and reliability are considered together for each decision.
- Every decision is accompanied by an explanation.
- Risky or uncertain decisions are not accepted automatically.
- Final decisions are recorded and can be audited later.
- The prototype depends on common tools: Python, SHAP, LIME, Fairlearn and cloud services.

7. Limitations

- The framework reduces the risks of AI decisions but cannot remove them.
- SHAP and LIME give approximate explanations, which can be misread.
- Findings on fairness and reliability depend on the quality and coverage of the dataset.
- The assessment layer adds computation and may raise the human-review rate; the size of both is still to be measured.
- Risk thresholds are set on a small number of development cases and may not transfer to other data.
- Only one model and a small dataset are planned, so results may not generalise.
- The prototype is not production-ready and is not intended for real confidential data.

VI. CONCLUSION AND FUTURE WORK

AI decisions made in the cloud can be biased, unreliable, opaque, or exposed to privacy and security risk, and the literature reviewed deals with explainability, trust, governance and frameworks largely in isolation [1]–[8]. This paper has proposed a framework that audits each decision once it is generated: it evaluates fairness, privacy, security and reliability, explains the output with SHAP or LIME, labels the risk, sends risky or uncertain cases to a person, and logs the final outcome for audit and monitoring. A small prototype using Python, SHAP, LIME, Fairlearn and a cloud platform is planned, to be compared against an unmodified baseline.

The framework does not claim to eliminate the risks of AI decisions. Since it has not yet been implemented or tested, no results are reported, and its usefulness can only be judged after the planned evaluation.

Future work includes:

- Testing other models, datasets and decision types.
- Adding and comparing further fairness, privacy and explanation methods.
- Tuning the risk score and thresholds on larger sets of decisions.

© 2026 IJSRET

- Aligning the framework with emerging standards such as IEEE P3396 [8].
- Building a simple monitoring interface that displays reviewer decisions and monitoring results.

REFERENCES

1. P. R. Kaza, "Explainable Superintelligence NLP with Agentic AI for Transparent Policy Decision-Making in Cloud Systems," 2026 IEEE International Conference on Emerging Computing and Intelligent Technologies (ICoECIT), 2026, doi: 10.1109/ICoECIT68303.2026.11497409.

2. Z. Wang, Y. Liu, and J. Huang, "An Open API Architecture to Discover the Trustworthy Explanation of Cloud AI Services," IEEE Transactions on Cloud Computing, 2024, doi: 10.1109/TCC.2024.3398609.
3. "Framework for Trustworthy AI/ML in B5G/6G," IEEE Conference Publication, 2022.
4. "AI-Governance and Levels of Automation for AIOps-supported System Administration," IEEE Conference Publication, 2020.
5. "Responsible AI, SDGs, and AI Governance in Africa," IEEE Conference Publication, 2022.
6. "Building Trust and Transparency in AI: A Review of Explainable AI and its Ethical Implications," IEEE Conference Publication, 2024.
7. "A Hybrid-DLT Based Trustworthy AI Framework," IEEE Conference Publication, 2024.
8. IEEE Computer Society, "IEEE Recommended Practice for Defining and Evaluating Artificial Intelligence (AI) Risk, Safety, Trustworthiness, and Responsibility," IEEE P3396.

© 2026 IJSRET