

Trustworthy AI: A Framework for Safe and Explainable Generative AI Systems

Author: **Tripti Sharma**

Department of Computer Science And Applications

IFTM University, Moradabad, Uttar Pradesh, India

Email: triptisharma7992@gmail.com

Abstract

Generative Artificial Intelligence (GenAI) systems have rapidly evolved from experimental technologies into widely used tools capable of generating text, computer programs, images, audio, video, recommendations, and other forms of digital content. Their increasing adoption in education, healthcare, business, cybersecurity, software development, and public services creates significant opportunities but also introduces new technical and societal risks. These risks include hallucination, bias, privacy leakage, security attacks, harmful content generation, prompt injection, unreliable reasoning, misuse, lack of transparency, and excessive dependence on automated outputs. Therefore, the development of trustworthy GenAI systems requires more than conventional accuracy testing.

This paper proposes a **Trustworthy Generative AI Lifecycle Framework (TGAI-LF)** for developing and evaluating safe and explainable generative AI systems. The proposed framework integrates seven major dimensions: governance and context, data and model assurance, safety and robustness, explainability and evidence, privacy and security, human oversight and accountability, and continuous monitoring. The framework is conceptually derived from established trustworthy-AI principles and risk-management guidance, particularly the NIST Artificial Intelligence Risk Management Framework (AI RMF), the NIST Generative Artificial Intelligence Profile, and the OECD AI Principles. The proposed approach connects common GenAI risks with preventive, detective, and corrective controls across the AI lifecycle.

Unlike approaches that treat explainability as an additional feature after model development, the proposed framework considers explainability, safety, privacy, security, and accountability as interconnected components of the complete AI system. An evaluation matrix is also proposed to assess trustworthiness using evidence-based indicators. The paper further discusses implementation challenges, trade-offs, limitations, and future research directions. The proposed framework can serve as a practical foundation for researchers and organizations seeking to design, evaluate, deploy, and govern safer and more explainable generative AI applications.

Keywords: Trustworthy AI, Generative AI, Explainable AI, AI Safety, AI Risk Management, Human Oversight, Privacy, Security, Accountability, Responsible AI

1. INTRODUCTION

Artificial Intelligence (AI) has become an important component of modern digital systems. Traditional AI applications have primarily focused on classification, prediction, recommendation, and decision-support tasks. In recent years, however, Generative Artificial Intelligence has introduced a new generation of systems capable of producing human-like text, images, audio, video, software code, and other forms of content.

Large Language Models (LLMs) and other foundation models have made GenAI accessible to users without specialized programming knowledge. Applications such as automated content generation, intelligent tutoring, code

generation, document summarization, customer support, cybersecurity assistance, and information retrieval are increasingly incorporating generative models.

Despite these benefits, GenAI systems introduce significant challenges. A generative system may produce an answer that appears highly convincing but is factually incorrect. It may reproduce biased information, disclose sensitive information, generate unsafe content, or respond unexpectedly to malicious prompts. When such systems are incorporated into organizational decision-making, these failures may result in financial, legal, social, privacy, or safety consequences.

Consequently, AI trustworthiness has become an important research area. The National Institute of Standards and Technology (NIST) describes trustworthy AI using several characteristics, including validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy enhancement, and fairness with harmful bias managed [1].

The NIST Generative AI Profile further identifies risks specifically associated with generative AI and provides recommendations for managing these risks throughout the AI lifecycle [2].

Similarly, the OECD AI Principles emphasize that AI systems should respect human rights and democratic values and should be transparent, explainable, robust, secure, safe, and accountable [3].

However, organizations still face a practical question:

How can these principles be converted into an integrated operational framework for a real-world generative AI application?

A model may have high benchmark accuracy but still be unsafe. Similarly, an AI system may generate explanations but fail to protect personal information. Therefore, trustworthy AI should not be treated as a single model characteristic.

This research proposes an integrated framework for safe and explainable GenAI systems by combining technical, organizational, and human-centered controls.

1.1 Problem Statement

Generative AI systems can fail in ways that are fundamentally different from conventional software systems.

Common problems include:

- hallucinated or fabricated information;
- biased or discriminatory outputs;
- generation of harmful content;
- privacy leakage;
- prompt injection;
- adversarial attacks;
- insecure tool usage;
- unauthorized data access;
- lack of meaningful explanations;
- excessive user trust;
- unclear responsibility for AI-generated decisions.

Traditional model evaluation often focuses on performance metrics such as accuracy, precision, recall, or benchmark scores. These metrics alone cannot determine whether a generative AI system is trustworthy.

Therefore, there is a need for an integrated framework that considers the complete AI lifecycle.

1.2 Research Gap

Existing trustworthy-AI standards provide important principles and risk-management functions. Explainable AI research provides techniques for interpreting model behavior. Security research provides methods for identifying adversarial attacks and vulnerabilities.

However, there remains an implementation gap between these individual areas.

In particular, organizations require a framework that connects:

1. risk identification;
2. model and data assurance;
3. safety;
4. explainability;
5. privacy;
6. security;
7. human oversight;
8. accountability; and
9. continuous monitoring.

The proposed research addresses this gap by integrating these dimensions into a single lifecycle-oriented framework for generative AI systems.

1.3 Research Objectives

The objectives of this research are:

1. To identify major trustworthiness requirements for generative AI systems.
 2. To study important risks associated with GenAI applications.
 3. To propose an integrated framework for safe and explainable GenAI.
 4. To integrate safety, robustness, explainability, privacy, security, fairness, and accountability.
 5. To propose an evaluation matrix for assessing GenAI trustworthiness.
 6. To identify implementation challenges and future research opportunities.
-

1.4 Research Questions

The research addresses the following questions:

RQ1: What characteristics are required for a generative AI system to be considered trustworthy?

RQ2: How can safety, explainability, privacy, security, and accountability be integrated into a single framework?

RQ3: How can the trustworthiness of a GenAI application be evaluated throughout its lifecycle?

RQ4: What challenges arise when implementing trustworthy AI principles in real-world GenAI systems?

2. BACKGROUND AND RELATED WORK

2.1 Trustworthy Artificial Intelligence

Trustworthy AI refers to AI systems that can be relied upon to operate appropriately within their intended context.

Trustworthiness is multidimensional. A system cannot be considered trustworthy simply because it produces accurate outputs.

For example, an AI model may correctly classify most inputs but:

- expose confidential information;
- discriminate against a particular group;
- fail under adversarial conditions;
- provide no explanation;
- encourage users to accept incorrect decisions.

NIST's AI RMF provides a structured approach for managing such risks through four core functions:

- **Govern**
- **Map**
- **Measure**
- **Manage**

These functions encourage organizations to understand AI risks, measure them, and implement appropriate management strategies [1].

2.2 Generative AI

Generative AI refers to AI systems capable of creating new content based on learned patterns.

Examples include:

- text generation;
- image generation;
- audio generation;
- video generation;
- code generation;
- synthetic data generation.

Generative AI differs from traditional predictive AI because the output space is often open-ended.

For example, a classification system may output:

"Spam"

or

"Not Spam"

A generative model may produce an entire paragraph explaining why an email is supposedly suspicious.

This flexibility creates additional uncertainty.

2.3 Risks Associated with Generative AI

Important GenAI risks include:

2.3.1 Hallucination

A hallucination occurs when a model generates information that appears plausible but is incorrect, unsupported, or fabricated.

For example, a model may invent:

- research papers;
- citations;
- statistics;
- legal cases;
- medical information;
- programming functions.

This is particularly dangerous in high-impact applications.

2.3.2 Bias

Generative models may reproduce biases present in training data or introduced during system development.

Bias may appear in:

- language;
- gender representation;
- cultural representation;
- occupation-related content;
- recommendations;
- decision support.

Therefore, fairness assessment should form part of trustworthy-AI evaluation.

2.3.3 Privacy Leakage

GenAI applications may process sensitive information including:

- names;
- email addresses;
- financial information;
- medical information;
- organizational documents;
- passwords or credentials.

Improper handling can result in privacy violations.

2.3.4 Prompt Injection

Prompt injection is an AI-specific security problem in which an attacker attempts to manipulate system behavior through specially designed instructions.

This becomes particularly important when GenAI applications have access to:

- databases;
 - APIs;
 - external tools;
 - files;
 - web resources;
 - organizational systems.
-

2.3.5 Misuse

Generative AI can potentially be misused for:

- misinformation;
- fraud;
- automated social engineering;
- malicious code generation;
- impersonation;
- harmful content generation.

Therefore, safety mechanisms should consider both accidental and intentional misuse.

3. EXPLAINABLE AI AND ITS IMPORTANCE

Explainable Artificial Intelligence (XAI) focuses on making AI behavior understandable to relevant users.

The importance of explainability depends on the application.

For example, an entertainment recommendation may require limited explanation, whereas an AI-assisted medical or financial recommendation may require much stronger justification.

A useful explanation can help users understand:

- what information influenced an output;
- what evidence supports the output;
- what limitations exist;
- how uncertain the system is;
- when human review is required.

However, explanation should not simply mean exposing technical model architecture.

An explanation should be **meaningful to its intended audience**.

For example:

Technical explanation

"Attention weights for tokens X and Y contributed to the prediction."

This may be useful for a researcher.

User-oriented explanation

"The recommendation was based mainly on the user's previous selections and the similarity of the current item to those selections."

This may be more useful to an ordinary user.

Therefore, the proposed framework treats explainability as **context-dependent evidence generation** rather than merely a model visualization technique.

4. RESEARCH METHODOLOGY

This research uses a **conceptual framework-development methodology**.

The methodology consists of four stages.

Stage 1: Identification of Trustworthiness Dimensions

Trustworthy-AI characteristics were identified from established AI risk-management principles and related research.

The major dimensions identified are:

- safety;
- reliability;
- robustness;
- explainability;
- privacy;

- security;
- fairness;
- accountability;
- human oversight.

Stage 2: Risk Categorization

Common GenAI risks were categorized into operational groups:

Risk Category Examples

Reliability	Hallucination, incorrect information
Safety	Harmful content
Security	Prompt injection, data exfiltration
Privacy	Sensitive-data leakage
Fairness	Bias and discrimination
Transparency	Lack of explanation
Accountability	Unclear responsibility
Human Factors	Over-reliance and automation bias

Stage 3: Control Mapping

Preventive, detective, and corrective controls were mapped to each risk.

Stage 4: Framework Development

The identified dimensions and controls were integrated into the proposed **Trustworthy Generative AI Lifecycle Framework (TGAI-LF)**.

This research does not claim experimental results. The framework is proposed for subsequent empirical validation.

5. PROPOSED TRUSTWORTHY GENERATIVE AI LIFECYCLE FRAMEWORK

The proposed framework consists of seven major layers.

Layer 1: Governance and Context

Layer 2: Data and Model Assurance

Layer 3: Safety and Robustness

Layer 4: Explainability and Evidence

Layer 5: Privacy and Security

Layer 6: Human Oversight and Accountability

Layer 7: Continuous Monitoring and Improvement

The layers interact throughout the AI lifecycle.

6. LAYER 1: GOVERNANCE AND CONTEXT

The first layer defines the purpose and boundaries of the AI system.

Before deployment, an organization should answer:

- What is the purpose of the AI system?
- Who will use it?
- Who may be affected?
- What decisions can it influence?
- What information will it access?
- What actions can it perform?
- What uses are prohibited?
- Who is responsible for the system?

A risk classification can be created.

For example:

Risk Level Example

Low	Creative writing assistant
Medium	Educational recommendation system
High	Financial decision support
Very High	AI influencing medical or safety-critical decisions

Higher-risk systems should receive stronger human review and testing.

7. LAYER 2: DATA AND MODEL ASSURANCE

Data quality has a direct effect on AI reliability.

The framework recommends:

7.1 Data Quality

Data should be evaluated for:

- accuracy;
- completeness;
- relevance;
- duplication;
- outdated information;
- bias.

7.2 Data Provenance

Organizations should maintain information about:

- source;
- collection method;
- processing;
- ownership;
- licensing;
- intended use.

7.3 Model Documentation

Model documentation should include:

- model name;
- version;
- provider;
- capabilities;
- limitations;
- intended use;
- known risks;
- evaluation results.

For third-party models, organizations should evaluate available provider documentation before integration.

8. LAYER 3: SAFETY AND ROBUSTNESS

Safety and robustness ensure that the system behaves appropriately under normal and abnormal conditions.

Testing should include:

Normal Inputs

Typical user requests.

Edge Cases

Unexpected but legitimate inputs.

Adversarial Inputs

Inputs designed to manipulate the system.

Misuse Scenarios

Requests that attempt to use the system for harmful purposes.

Failure Scenarios

Cases where the model does not have sufficient information.

A trustworthy system should not always attempt to produce an answer.

Sometimes the safest response is:

"I do not have sufficient reliable information to answer this request."

This is an important component of trustworthy behavior.

9. LAYER 4: EXPLAINABILITY AND EVIDENCE

The proposed framework introduces an **evidence-based explainability layer**.

The system should provide appropriate supporting information where feasible.

Possible evidence includes:

- source references;
- retrieved documents;
- input factors;
- confidence or uncertainty indicators;
- limitations;
- reasoning summaries;
- human-review status.

For example, instead of simply generating:

"This information is correct."

the system could provide:

"This response is based on the following retrieved sources. The system has moderate confidence because the available sources do not fully agree."

This provides a more meaningful representation of uncertainty.

10. LAYER 5: PRIVACY AND SECURITY

Privacy and security are essential components of trustworthy GenAI.

10.1 Privacy Controls

Recommended controls include:

- data minimization;
- access control;

- encryption;
- secure storage;
- retention policies;
- anonymization where appropriate;
- restricted access to sensitive prompts;
- controlled data transmission.

10.2 Security Controls

Security testing should consider:

- prompt injection;
- unauthorized access;
- data exfiltration;
- malicious files;
- insecure APIs;
- unsafe tool calls;
- compromised dependencies.

When GenAI is connected to external tools, the risk increases because the model may move from simply generating information to taking actions.

Therefore:

AI-generated instructions should not automatically be treated as authorized actions.

Appropriate authorization and validation should occur before consequential actions.

11. LAYER 6: HUMAN OVERSIGHT AND ACCOUNTABILITY

Human oversight is particularly important for high-impact AI systems.

The framework recommends three levels.

Level 1: Human-in-the-Loop

A human reviews the AI output before an important action is taken.

Level 2: Human-on-the-Loop

The system operates automatically but humans continuously monitor it and can intervene.

Level 3: Human-in-Control

Humans retain ultimate authority over important decisions.

The appropriate level depends on the application risk.

11.1 Accountability

Accountability requires clear responsibility.

Organizations should identify:

- AI system owner;
- technical developer;
- data owner;
- security responsible person;
- human reviewer;
- decision authority.

Important system events should also be logged where legally and technically appropriate.

12. LAYER 7: CONTINUOUS MONITORING

Trustworthiness is not permanent.

A model may become less reliable because of:

- model updates;
- data changes;
- user behavior;
- new attack methods;
- changes in external information;
- changes in system integrations.

Therefore, monitoring should continue after deployment.

Important monitoring indicators include:

- hallucination incidents;
- unsafe outputs;
- security incidents;
- privacy incidents;
- user complaints;
- explanation failures;
- model performance degradation;
- unusual usage patterns.

When serious problems are detected, the organization should be able to:

1. investigate;
2. restrict the system;
3. modify controls;
4. retrain or replace the model;
5. roll back an update;
6. temporarily suspend the application.

13. FRAMEWORK ARCHITECTURE

The proposed architecture can be represented as:

Governance & Context



Data and Model Assurance



Safety & Robustness



Explainability & Evidence



Privacy & Security



Human Oversight & Accountability



Continuous Monitoring



Feedback to Governance

This should be treated as a **continuous lifecycle**, not a one-time sequence.

Monitoring results can cause the organization to return to an earlier layer.

For example:

Monitoring → Security problem detected → Risk reassessment → New security controls → Retesting → Deployment

14. RISK-CONTROL MAPPING

Risk	Preventive Control	Detective Control	Corrective Control
Hallucination	Retrieval, validation	Fact checking	Human correction
Bias	Data assessment	Fairness testing	Dataset/model modification

Risk	Preventive Control	Detective Control	Corrective Control
Privacy leakage	Data minimization	Privacy monitoring	Access restriction
Prompt injection	Input filtering	Attack testing	Session termination
Harmful output	Safety policies	Red teaming	Blocking/escalation
Incorrect decision	Human approval	Decision audit	Override
Security attack	Authentication	Security monitoring	Incident response
Lack of explanation	Explanation design	Explanation evaluation	Human review
Model drift	Controlled updates	Continuous monitoring	Reassessment

15. PROPOSED EVALUATION MATRIX

A single score cannot fully represent AI trustworthiness.

Therefore, the proposed framework uses multiple dimensions.

A five-level assessment scale can be used:

- **0 – Not addressed**
- **1 – Initial**
- **2 – Partially implemented**
- **3 – Implemented**
- **4 – Continuously monitored and improved**

Dimension	Possible Metric	Evidence
Safety	Unsafe-output rate	Red-team report
Reliability	Factual accuracy	Benchmark results
Explainability	Explanation usefulness	User/expert evaluation
Privacy	Data exposure incidents	Privacy audit
Security	Attack success rate	Security testing
Fairness	Group disparity	Fairness evaluation
Accountability	Traceability	Audit logs
Human Oversight	Override effectiveness	Review records
Monitoring	Detection/response time	Operational logs

The purpose of this matrix is not to create an arbitrary universal "trust score." Rather, it provides structured evidence for decision-making.

16. DISCUSSION

The proposed framework demonstrates that trustworthy GenAI requires integration of multiple dimensions.

A system may be technically accurate but still fail to be trustworthy.

For example, consider an AI educational assistant.

Suppose the system provides correct answers in most cases. However:

- it collects unnecessary student information;
- it cannot explain important recommendations;
- students become excessively dependent on it;
- the system generates biased content;
- teachers cannot review important outputs.

Such a system would not be fully trustworthy despite good accuracy.

Therefore, trustworthy AI should be understood as a **socio-technical property**.

16.1 Explainability versus Privacy

One important challenge is the relationship between explainability and privacy.

Providing more information may improve transparency, but it may also reveal:

- personal information;
- confidential data;
- proprietary information;
- security-sensitive details.

Therefore, explanations should provide useful evidence without unnecessarily exposing sensitive information.

16.2 Safety versus Usability

Overly strict safety controls can make an AI system difficult to use.

For example, a safety filter might incorrectly block legitimate educational or cybersecurity questions.

Therefore, safety mechanisms should be evaluated for both:

- harmful-content prevention;
- legitimate-use preservation.

This is an important reason why testing should include both misuse and normal-use scenarios.

16.3 Automation Bias

Users may assume that AI-generated content is correct because it appears fluent and professional.

This phenomenon can lead to automation bias.

Therefore, systems should communicate uncertainty and encourage human verification when appropriate.

For high-risk applications, human review should remain mandatory.

16.4 Accountability Across Multiple Organizations

Modern AI applications can involve:

- foundation-model provider;
- application developer;
- cloud provider;
- data provider;
- tool provider;
- organization deploying the application;
- final human user.

When something goes wrong, determining responsibility can therefore become difficult.

The proposed framework addresses this through explicit role definition and documentation.

17. PRACTICAL IMPLEMENTATION ROADMAP

Organizations can implement the proposed framework using the following steps.

Step 1: Define the Use Case

Clearly define:

- purpose;
- users;
- affected individuals;
- intended outputs;
- prohibited uses.

Step 2: Conduct Risk Assessment

Identify:

- safety risks;
- privacy risks;
- security risks;
- fairness risks;
- reliability risks.

Step 3: Document Data and Model

Record:

- model version;
- data sources;
- limitations;
- dependencies;

- ownership.

Step 4: Perform Pre-Deployment Testing

Test:

- normal cases;
- edge cases;
- adversarial cases;
- misuse cases.

Step 5: Implement Human Oversight

Define when:

- human approval is required;
- escalation occurs;
- AI output can be rejected;
- AI actions can be overridden.

Step 6: Implement Monitoring

Monitor:

- incidents;
- security events;
- privacy events;
- user feedback;
- performance;
- unsafe outputs.

Step 7: Periodic Reassessment

The complete system should be reassessed after:

- major model updates;
- major data changes;
- new integrations;
- significant incidents;
- regulatory changes.

18. CHALLENGES AND LIMITATIONS

The proposed framework has several limitations.

18.1 Conceptual Nature

The framework has not yet been validated through a large-scale experimental study.

Therefore, its effectiveness cannot be claimed as an empirical result.

18.2 Domain Dependency

Trustworthiness requirements differ between domains.

For example, a healthcare AI system requires stronger safeguards than a simple creative-writing application.

18.3 Measurement Difficulty

Some characteristics, such as trust, fairness, explainability, and accountability, are difficult to represent using one numerical value.

18.4 Resource Requirements

Continuous monitoring and red-team testing may require:

- technical expertise;
- financial resources;
- infrastructure;
- trained reviewers.

18.5 Rapidly Changing Technology

GenAI technology evolves quickly. New models, tools, attacks, and risks may appear after a framework has been implemented.

Therefore, the framework itself should remain adaptable.

19. FUTURE WORK

Future research can extend this work in several directions.

First, the proposed framework can be empirically evaluated using real-world GenAI systems.

Second, researchers can develop validated quantitative metrics for each trustworthiness dimension.

Third, user studies can determine whether explanations actually improve:

- user understanding;
- decision quality;
- appropriate trust;
- error detection.

Fourth, automated red-teaming methods can be integrated with continuous monitoring.

Fifth, future research should examine trustworthy AI in **agentic AI systems**, where models can independently interact with tools and external systems.

Sixth, domain-specific versions of the framework can be developed for:

- education;
- healthcare;

- finance;
- cybersecurity;
- public administration.

Finally, future studies can examine the relationship between explainability and automation bias to determine whether explanations actually reduce inappropriate reliance on AI.

20. CONCLUSION

Generative Artificial Intelligence has significant potential to transform education, healthcare, business, software engineering, cybersecurity, and many other domains. However, its benefits cannot be fully realized unless the associated risks are systematically managed.

This paper proposed the **Trustworthy Generative AI Lifecycle Framework (TGAI-LF)** for designing and evaluating safe and explainable generative AI systems.

The framework integrates seven major dimensions:

1. Governance and Context
2. Data and Model Assurance
3. Safety and Robustness
4. Explainability and Evidence
5. Privacy and Security
6. Human Oversight and Accountability
7. Continuous Monitoring and Improvement

The central contribution of the framework is its lifecycle-oriented approach. Trustworthiness is not treated as a single property of a model or as an explanation feature added after development. Instead, it is considered a property of the complete socio-technical AI system.

The framework also provides an evaluation matrix through which organizations can collect evidence about safety, reliability, explainability, privacy, security, fairness, accountability, human oversight, and monitoring.

Because this study is conceptual, future empirical research is necessary to validate the framework through case studies, user studies, security testing, and quantitative evaluation.

Ultimately, trustworthy GenAI should not merely be capable of generating useful content. It should also be **safe, explainable, secure, privacy-aware, accountable, robust, and subject to meaningful human oversight**.

REFERENCES

- [1] E. Tabassi, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, USA, Jan. 2023. doi: 10.6028/NIST.AI.100-1.
- [2] C. Autio, R. Schwartz, J. Dunietz, S. Jain, M. Stanley, E. Tabassi, P. Hall, and K. Roberts, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," NIST AI 600-1, National Institute of Standards and Technology, Gaithersburg, MD, USA, Jul. 2024. doi: 10.6028/NIST.AI.600-1.

[3] OECD, “OECD AI Principles,” OECD.AI, updated 2024.

[4] OECD, *Advancing Accountability in AI: Governing and Managing Risks Throughout the Lifecycle for Trustworthy AI*, OECD Publishing, 2023. doi: 10.1787/2448f04b-en.

[5] Google DeepMind, “Holistic Safety and Responsibility Evaluation of Generative AI Systems,” arXiv:2404.14068, 2024.

[6] NIST, “AI Risk Management Framework Resources,” National Institute of Standards and Technology.

[7] OECD, *Governing with Artificial Intelligence*, OECD Publishing, 2024.
