



Artificial Intelligence Working Group

Blockchain Maturity Model (BMM)
Artificial Intelligence (AI) Supplement

Version: 1.0

Last Modified: September 22, 2026

DRAFT

Yeshwant Muthusamy, PhD	AI Working Group Leader Title	TBD Date
Meiyappan Masilamani	Dir., Standards & Certification Title	TBD Date



Table of Contents

Chair's v0.8 Change Summary	iii
1 Introduction	1
1.1 Purpose	1
1.2 Scope	2
1.3 Use	2
1.4 General Challenges	3
1.5 Securing AI Systems: Specific Challenges	5
2 Role of Blockchain in Artificial Intelligence	6
3 AI Security Assessment Subject Matter Expert (AI SEC SME) Qualifications	9
3.1. Overview	10
3.2. Recommended Training Components	10
3.3. Qualification Requirements	12
4 Solution Requirements	13
4.1 Requirement Language	14
4.2 Source Code Availability and Transparency	15
4.3 AI Model Availability and Versioning	19
4.4 Open-Source Licensing, Interoperability, and Process Transparency	22
4.5 Dataset Availability and GDPR Compliance	24
4.6 Training Loop Availability	29
4.7 AI Model Weights and Artifacts	40
4.8 All Actors' Signatures and Accountability	43
4.9 Model Output Visibility and Real-Time Monitoring	48
4.10 Additional Requirements for High-Risk and Critical-Infrastructure Solutions	55
5 Assessment Methodology and Compliance Determination	63
5.1 Overview	63
5.2 Assessment Team Composition	64
5.3 Evidence Collection and Review	65
5.4 Risk-Type Determination	66



GBA

BMM Artificial Intelligence Supplement

5.5	Compliance Determination	68
5.6	AI Supplement Rating and Certification Level	70
5.7	Assessment Report and Recommendations	72
Appendix A: Glossary		73
Appendix B: Mapping to Global AI Frameworks		77
Appendix C: [INFORMATIVE] Quantum Readiness Overview		80
	Purpose	80
	Why it matters	80
	Assessor focus	80
	Further guidance	81
Bibliography		82



Chair's v0.8 Change Summary

This table records the Chair's changes after the AI Working Group's one-month review period. It is included for transparency and does not replace the change-control history. The changes are intended to improve assessments, reduce duplication, and keep this supplement focused on its purpose.

Area	v0.7	v0.8 change	Rationale
Version and governance record	Version 0.7; last modified August 25, 2026.	Updated to Version 0.8; last modified September 14, 2026; added Chair's v0.8 change summary and v0.8 change-control entry.	Makes the Chair-directed revision transparent after the Working Group review period.
AI SEC appendices	Appendix C contained a training AI SEC model; Appendix D contained a detailed AI SEC API example.	Removed both appendices from this supplement.	Section 3 already defines assessor competencies and qualifications. The removed material is non-normative, duplicates or exceeds the supplement's purpose, and is better maintained as separate training or implementation guidance.
Quantum material	Standalone Section 6 contained seven subsections, technical background, standards tables, and eight quantum readiness requirements.	Replaced Section 6 with Appendix C: [INFORMATIVE] Quantum Readiness Overview. Removed quantum computing-related terms from the Glossary in Appendix A.	Keeps assessors aware of the quantum risk while avoiding a separate technical quantum-assessment regime. Operative controls remain in Section 4.
Quantum requirements	Section 6.6 introduced separate quantum readiness requirements and risk-type applicability.	Removed standalone Section 6.6 requirements; Appendix C points assessors to cryptographic agility, Section 4.8 controls, lifecycle planning, and incident response.	Avoids duplicate controls and specialized technical detail not required for BMM AI assessors. Maintains the BMM Quantum Computing Supplement the authoritative source



			for all information on quantum computing.
Artifact provenance	Parallel on-chain hash/provenance language appeared in 4.2.4, 4.3.7, 4.5.7, 4.7.1, and 4.6.20.	Retained artifact-specific evidence but aligned the related rows to a common Artifact Provenance and Lifecycle Control centered on 4.6.20.	Reduces duplicate requirements while preserving evidence for code, datasets, models, model weights, and training runs.
Lifecycle change control	Separate change-log requirements appeared in 4.2.9, 4.3.11, and 4.5.8.	Consolidated the common lifecycle change-control elements in 4.3.11; source-code and dataset rows now cross-reference it.	Creates one coherent audit trail for changes while retaining artifact-specific detail.
Reproducibility	Seeds, environment provenance, and replication evidence were distributed across 4.6.4, 4.6.13, 4.6.18, and 4.6.21.	Consolidated the reproducibility package in 4.6.18; related rows now identify required evidence elements.	Makes the assessor test one reproducibility control rather than several overlapping controls.
Monitoring and alerts	Performance monitoring, drift detection, alert recording, escalation, and resolution were divided across 4.9.8 and 4.9.10–4.9.14.	Consolidated production performance/drift monitoring in 4.9.8 and alert-to-resolution workflow in 4.9.12; supporting rows cross-reference them.	Reflects how these controls operate in practice and avoids fragmented assessment findings.
Human oversight	Human-review scope, thresholds, overrides, and reviewer qualifications were separate controls in 4.10.4–4.10.7.	Consolidated them under meaningful human oversight in 4.10.4; supporting rows retain the associated evidence elements and applicability.	Clarifies that meaningful oversight is one integrated control.
Incident management	Incident response, post-incident review, and remediation timelines were separate controls in 4.10.16–4.10.18.	Consolidated the lifecycle under 4.10.16; supporting rows retain investigation and remediation evidence.	Ensures assessors evaluate the entire incident lifecycle.
Risk-type consistency	Narrative summaries could be read as	Added language making the Section 4 applicability	Prevents a summary statement from



	conflicting with applicability cells in Section 4.	tables controlling in the event of conflict. Revamped the section on “Risk Type Implications for Assessment” in Section 5.4 to remove inconsistencies with respect to Section 4 requirements. The previous prose was making blanket assertions that were inconsistent with the latest revisions to Sections 4.2 and 4.4.	overriding requirement-level designations.
Assessment team	Section 5.2 required at least four members but listed three roles.	Added the Lead Assessor as the fourth defined team role.	Resolves an internal inconsistency and clarifies accountability for final determinations.
Alert terminology	Section 4.9.13 used “Type 1/2/3 alert,” overlapping with solution risk types.	Changed alert labels to Severity 1, Severity 2, and Severity 3.	Prevents confusion between alert severity and solution risk type.
Cross-references and editorial cleanup	Several internal labels and numbering references were inconsistent.	Corrected selected Section 4.9 and 4.10 cross-references, Appendix B mapping labels, AI category numbering, and related editorial inconsistencies.	Improves usability and reduces avoidable assessment ambiguity.

Change Control History:

Version	Date	Author(s)	Summary
0.1	Sep 26, 2023	Mitja Goroshevsky	Initial draft with input from the GBA AI Working Group.
0.2	Jan 10, 2024	Gerard Dache	Restructured, formatted and distributed as a draft with request for comments.
0.3	Jan 19, 2026	Yeshwant Muthusamy	Formatting, typo fixes and addition of new content, in preparation for the AI Working Group to resume operations and make further updates (if necessary).



0.4 WORKING DRAFT	June 23rd, 2026	Yeshwant Muthusamy	Began major structural and content changes for review by the WG. Version number will not change until all changes are in, but the date will change to reflect the latest changes. Changes that have been reviewed and discussed in WG calls will have the old prose deleted in change tracking.
0.5	June 23, 2026	Yeshwant Muthusamy	First Baseline DRAFT incorporating all revisions. Ready for AI WG review. Please comment directly on the document in NextCloud for ease of tracking all comments and revisions.
0.6	August 18, 2026	Yeshwant Muthusamy	AI Working Group review period CLOSED; incorporated (most) review comments. Will be taken off Google Docs soon for final formatting and cleanup.
0.7	August 25, 2026	Yeshwant Muthusamy	Moved to Microsoft Word; 0.6 marked READ-ONLY on Google Docs. All further changes will be made in Word.
0.8	September 14, 2026	Yeshwant Muthusamy	Chair's consistency and scope revision following the Working Group review period. Removed standalone AI SEC training and API appendices; condensed quantum material into an informative appendix; streamlined overlapping requirements; corrected cross-references, terminology, numbering, and formatting inconsistencies. Removed quantum computing-related terms from the Glossary in Appendix A.
0.9	September 15, 2026	Yeshwant Muthusamy	Version uploaded to Google Docs after removing Appendix D, for final round of comments.
1.0	September 22, 2026	Yeshwant Muthusamy	Two out of three late-breaking comments have been incorporated into this version that might benefit from wider review. New section 4.1.2 serves as a bridge to a comprehensive architecture-neutral rewrite of Section 4's ML requirements, which is deferred to Version 1.1 and will be assigned to the commenter. Sections 5.5.1 – 5.5.3 on Partial vs. Non-



GBA

BMM Artificial Intelligence Supplement

			Compliant decision rules clarifies the points made in Section 5.5. The third comment, on making the risk classification methodology is completely deferred to Version 1.1 as it merits a more involved WG discussion before we can agree on a way forward. Providing the initial prose for WG discussion will also be assigned to the commenter. Other edits include minor typo and formatting fixes. This is the first version released for general GBA feedback.
--	--	--	---



1 Introduction

Artificial intelligence (AI) refers to machine-based systems that infer from input data how to generate outputs, such as predictions, content, recommendations, or decisions, that influence physical or virtual environments. AI systems can vary in autonomy and adaptiveness, including systems that self-update, operate with greater independence, or evolve after deployment [5].

This supplement extends the Blockchain Maturity Model (BMM) framework [7] to address four categories of AI relevant to blockchain systems:

1. Discriminative AI, which classifies data into categories (e.g., speech recognition, image recognition, sentiment analysis)
2. Predictive AI, which generates predictions from historical data (e.g., risk prediction, trend forecasting)
2. Generative AI, which produces new content (text, images, video) or code
3. Agentic AI, which involves one or more agents acting with a degree of autonomy to pursue objectives

We define an AI system as an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy [2]. This definition acknowledges that AI systems include technical components, human oversight and governance, operational context, deployment environments, and software.

The BMM AI supplement applies to solutions where blockchain technology is used to enhance the trustworthiness, accountability, and auditability of AI systems.

1.1 Purpose

This document describes the AI-specific governance, risk management, and transparency requirements that solutions integrating AI with blockchain technology must satisfy to achieve a BMM AI Trusted Solution rating at specified BMM maturity levels [7]. The supplement is applicable to solutions where blockchain technology is used to enhance the trustworthiness, accountability, and auditability of AI systems, complementing baseline BMM assessments and aligning with international frameworks and standards, including:

- ISO/IEC 42001:2023 (AI Management Systems) [4]
- OECD AI Principles [5]
- EU Artificial Intelligence Act (Regulation 2024/1689) [3]

Commented [1]: Alternate opening:

Artificial intelligence (AI) capabilities have expanded significantly since 2021. At a high level, AI refers to machine-based systems that infer from input data how to generate outputs, such as predictions, content, recommendations, or decisions, that influence physical or virtual environments. In 2019, the OECD published an initial definition focused on predictive and analytical AI systems. The OECD AI Principles were updated in 2024 to reflect that AI systems can vary in autonomy and adaptiveness, including systems that self-update, operate with greater independence, or evolve after deployment.

This supplement extends the Blockchain Maturity Model (BMM) framework to address three categories of AI relevant to blockchain systems:

1. Predictive/analytical AI, which generates forecasts or classifications from historical data
2. Generative AI, which produces new content or code
3. Agentic AI, which acts with a degree of autonomy to pursue objectives

The BMM AI supplement applies to solutions where blockchain technology is used to enhance the trustworthiness, accountability, and auditability of AI systems.

It describes the AI-specific governance, risk management, and transparency requirements that such solutions must satisfy to achieve a BMM AI Trusted Solution certification, and aligns with major international standards and regulations, including ...

Commented [2]: @minovac@web3strategygroup.com I am OK with most of it, except for specifically calling out OECD in the first paragraph. OECD is one of many standards we are looking to play nicely with, and calling it out so early in the doc could be perceived unfavorably by some members of GBA. I will incorporate the rest of your suggestion into the prose.

Commented [3]: Why not mention OECD ? They were one of first to define AI. Does not detract from our paper.

Commented [4]: We can, but just not open with it. That was my point.

Commented [5]: This revision clarifies the scope and positions the supplement within the global AI governance ecosystem. It establishes explicit alignment with major international standards and regulations rather than operating in isolation. The language shifts from "BMM Trusted AI Solution" to "BMM AI Trusted Solution certification" to indicate a formal assessment pathway.



- NIST AI RMF 1.0 [2]

Appendix B goes into detail on how the various requirement sub-sections in Section 4 (Solution Requirements) map to relevant clauses in these international frameworks and standards.

1.2 Scope

The AI Supplement applies to solutions where AI systems generate critical outputs, decisions, or recommendations that materially affect stakeholders, and where blockchain or distributed ledger technology enhances the verifiability, auditability, or integrity of those AI-driven decisions.

Scope Classification (Risk-Based Types):

The BMM AI risk types are informed by international risk-based approaches, including the EU AI Act [3], but constitute an independent assessment taxonomy.

Type 1 (Low-Risk AI): Pattern recognition, optimization recommendations, non-critical analytics (e.g., content recommendations, inventory optimization, sentiment analysis).

Type 2 (High-Risk AI): Decisions affecting individual rights, financial transactions, hiring/lending, content moderation, or access to essential services.

- Data protection (GDPR/privacy) and bias mitigation are required.

Type 3 (Critical Infrastructure AI): Autonomous systems, healthcare diagnostics, safety-critical decision-making, military/national security applications, or systems where failure could cause loss of life or significant harm.

- Enhanced post-deployment monitoring, real-time audit trails, human-in-the-loop validation, and incident reporting protocols are mandatory ("shall").
- For Type 3 systems deployed in disconnected, sovereign, or tactical environments, these mandatory requirements shall extend to the physical infrastructure layer, including hardware supply chain integrity, modular enclaves, and localized data-to-compute co-location.

1.3 Use

This AI supplement along with the BMM document is designed to be used as a guide both for solution implementers and assessors of the solution. To be rated at BMM Level X for AI, a solution must:

Commented [6]: Suggest:

"The BMM AI Trusted Solution supplement describes artificial intelligence (AI) governance, risk management, and transparency requirements for solutions that integrate AI with blockchain. The supplement complements baseline BMM assessments and aligns with international AI frameworks and standards"

This aligns with the other 9 BMM Supplements. They do not discuss specific BMM ratings or maturity levels.

Commented [7]: Add: AI System Definition

The BMM AI supplement, in line with NIST AI RMF, defines an AI system as an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

This definition acknowledges that AI systems include ...

Commented [8]: Incorporated this AI system definition into the Introduction section as it lays the groundwork for the rest of the document. Also, the last paragraph of your Introduction is similar to the first sentence of the Purpose subsection, so deleted it and incorporated into the Purpose section.

Commented [9]: The 2021 date was not supported by any reference, so removed it. AI has been a field of active research since 1950 ("Can Machines Think?" paper by Alan Turing started it all) and several landmark/watershed inventions were done way before 2021, so I was not sure what was significant about ...

Commented [10]: This introduces a risk-tiered approach matching the EU AI Act methodology. The three-tier system acknowledges that not all AI solutions require the same rigor—low-risk systems face lighter requirements while critical infrastructure demands maximum scrutiny. This makes the supplement ...

Commented [11]: ? EU AI Act categorizes AI systems into 4 levels of risk: unacceptable risk (prohibited), high risk (strict compliance obligations before market entry), limited risk (transparency disclosures to users) and minimal risk (no mandatory obligations)

Commented [12]: The current phrase "aligned with the EU AI Act's risk-based approach" may create the impression that the three BMM types correspond directly to the legal categories established by the EU AI Act.

Commented [13]: Alex, I have removed the second sentence containing "legal classification" for obvious reasons.



- Satisfy baseline BMM element requirements at Level X (all 11 elements), AND
- Satisfy all AI Supplement requirements applicable to the solution's risk type (as defined above), AND
- Demonstrate compliance with applicable laws and regulations in the solution's operating jurisdiction(s).

1.4 General Challenges

The proliferation of AI systems across industries and sectors has created both opportunities and risks. While AI-enabled systems offer significant benefits, including improved efficiency, data-driven decision-making, and innovation, they also present novel security, privacy, and ethical risks that traditional software governance frameworks often fail to address.

Blockchain and distributed ledger technologies can enhance accountability and auditability by providing:

- Immutable audit trails of AI model training data, model updates, and decision-making processes
- Transparent versioning and change-tracking across the AI system lifecycle
- Decentralized verification of AI system integrity and compliance
- Cryptographic proof of model provenance and training provenance

The AI Supplement establishes governance requirements to ensure that blockchain-enabled AI systems operate responsibly, transparently, and in compliance with applicable legal and ethical standards.

AI governance presents operational challenges that this supplement addresses through blockchain-enabled auditability and transparency. These challenges are organized into five categories:

A. Data and Model Integrity

- **Data Poisoning:** Manipulation of training data to cause incorrect AI model behavior (e.g., injecting mislabeled samples to degrade model accuracy)
- **Model Inversion & Membership Inference:** Unauthorized extraction of sensitive information from models or training sets through adversarial queries
- **Dataset Drift:** Undetected distribution shifts in live data causing model performance degradation over time
- **Hardware and Environment Divergence:** Failure to replicate exact physical training environments (e.g., specific GPU types, host driver variants, and library versions like

Commented [14]: The original "democratization" framing was abstract and didn't clearly connect to blockchain's role. This revision focuses on the practical value proposition: why blockchain matters for AI governance. It positions the supplement as enabling responsible innovation rather than restricting AI development. The bullet points make the benefits concrete and testable.]

Commented [15]: So long as you are implementing key Web3 services / tech that will allow for this democratization and scalability for 'DeVerification' of AI services, you can state this; in particular, DID (Decentralized Identity) services for every citizen / consumer that also sits on a government body / AI institution-led infrastructure that governs the entire use of AI, based on the rules or regulations / data privacy laws put in place by society. This means having a 100% customizable Rules Engine-as-a-Service, that allows for you to enter the data governance details (GDPR, CCPA, etc.) as text, convert it to code, then have it codified into every DID as rights / protections for the citizens. If you don't have that, then there's no point in trying to keep up with the pace of AI. This is possible now, see the examples below in terms of those as-a-Service ideas, as well as what Web2 vs Web3 use in the AI space will eventually mean for us all:

DID and Rules Engine as-a-Service (with History of the Web Intro):
<https://socialchangeplatforms.com/wp-content/uploads/2023/09/Web1-to-Web3-History-Our-Web3-Products.pdf>

Web2 vs Web3 Summary w/ AI:
<https://socialchangeplatforms.com/wp-content/uploads/2025/12/Tabor-Speaking-Event-AI-Summary-Web2-vs-Web3.pdf>

Commented [16]: Decentralized Identity not in scope for BMM.



CUDA/ROCm). Without infrastructure provenance logs alongside dataset hashes, deterministic model replay and cryptographic verification of model lineage become impossible.

B. Robustness and Safety

- **Adversarial Attacks:** Crafted inputs designed to fool AI systems (e.g., perturbations that cause misclassification)
- **Out-of-Distribution Generalization:** Poor model performance on data unlike training examples, leading to unreliable predictions
- **Autonomous System Failures:** Unintended behaviors in self-governing or highly autonomous AI systems
- **Physical and Silicon-Layer Exploits:** Vulnerabilities within the underlying compute hardware supply chain, including compromised device firmware (BIOS, BMC, NIC, or GPU firmware). Additionally, Type 3 systems operating at the tactical edge face severe physical security, thermal, and power constraints that can trigger unhandled autonomous system failures if not managed by ruggedized, modular architectures.

C. Privacy and Data Protection

- **Unauthorized Data Access:** Breaches exposing training data or personal information contained in models
- **GDPR Right-to-Erasure Conflicts:** Tension between blockchain immutability and data subject rights to deletion (addressed in Section 4.5)
- **Differential Privacy & Federated Learning:** Techniques to protect individual privacy while maintaining model utility
- **Data-in-Transit Vulnerabilities:** Exposure of sensitive or sovereign datasets as they move between disparate storage networks and centralized hyperscaler compute nodes. Mitigating this requires evaluating the physical co-location of data and compute utilizing ultra-high-speed bus interconnects and GPU-Direct Storage (GDS) to prevent mid-transit interception.

D. Transparency and Accountability

- **Black-Box Decision-Making:** Lack of explainability in AI system outputs, making auditing and accountability difficult
- **Bias and Discrimination:** Systematic unfairness in AI system decisions affecting protected groups (gender, race, age, disability)
- **Model Attribution:** Unclear responsibility for AI model failures or unintended outcomes

E. Governance and Compliance



- **Regulatory Fragmentation:** Inconsistent AI regulations across jurisdictions (EU, US, China, etc.)
- **Audit and Forensic Capabilities:** Difficulty tracing AI system changes and decisions post-incident
- **Lack of Standardized AI Assessment Expertise:** Shortage of assessors qualified to evaluate AI systems against objective criteria
- **Infrastructure Sustainability and Resource Isolation:** The compounding environmental impact (Power Usage Effectiveness (PUE) and cooling metrics) of large-scale AI clusters. Operationally, this includes a lack of auditable telemetry within orchestration layers (e.g., Kubernetes, Slurm) to verify strict multi-tenant resource isolation and secure hardware decommissioning/data-wiping protocols.

1.5 Securing AI Systems: Specific Challenges

1.5.1. Technical Security Risks:

- **Model Compromise:** Unauthorized modification of model weights or architecture (e.g., backdoor insertion)
- **Training Pipeline Vulnerability:** Attacks on data collection, preprocessing, or feature engineering stages
- **Inference-Time Attacks:** Adversarial inputs, prompt injection (for LLMs), or behavioral manipulation
- **Supply Chain Vulnerabilities:** Malicious dependencies in libraries, frameworks, or datasets (e.g., compromised PyTorch package)

1.5.2. Operational and Governance Risks:

- **Inadequate Change Management:** Insufficient controls over model updates and retraining cycles
- **Insufficient Human Oversight:** Automated decisions lacking meaningful human review or intervention capability
- **Post-Deployment Drift:** Model performance degradation in production without detection or alerting
- **Incident Response Gaps:** Inability to investigate or remediate AI system failures due to incomplete audit trails

Commented [17]: Could do with some word-smithing to elaborate on the actual results of that risk. Make it more consistent with the prose in 1.5.1.



The requirements in Section 4 (Solution Requirements) are designed to mitigate these risks by establishing:

- **Immutable audit trails** for traceability (all lifecycle changes recorded on blockchain)
- **Cryptographic signatures** for accountability (all actions attributed to authorized actors)
- **Decentralized oversight mechanisms** for verification (third-party auditing without single point of trust)
- **Explicit model versioning** and change logs for forensic analysis post-incident

2 Role of Blockchain in Artificial Intelligence

This section explains how blockchain technology enhances AI governance. Blockchain provides six key capabilities that address the challenges outlined in Section 1.

2.1. Transparency: Artifact Availability and Accessibility

Definition: Stakeholders can access AI system documentation, code, models, training data metadata, and performance metrics necessary to understand how the AI system works and make informed decisions about its use.

Blockchain Role: Blockchain-backed decentralized storage (e.g., IPFS, Arweave, Filecoin) ensures that artifacts remain accessible and verifiable over the AI system's lifetime. Hash commitments recorded on-chain provide cryptographic proof of artifact integrity and timestamp, preventing tampering or unauthorized modification.

Examples:

- Source code published on decentralized repository (GitHub, GitLab) with version control history; blockchain records hash of each release
- Model architecture and training dataset metadata published via IPFS with on-chain registration transaction
- Performance metrics and validation test results recorded with timestamp and authorized accessor signatures

2.2. Visibility: Lifecycle Event Recording

Definition: Changes to AI systems, training data, model parameters, or deployment configurations are recorded in real-time and visible to authorized stakeholders for monitoring and investigation.

Blockchain Role: Smart contracts or transaction logs record every significant AI lifecycle event (model retraining, dataset update, production deployment, performance threshold breach). Immutable recording enables post-incident forensic analysis and compliance verification.

Commented [18]: The original bullet-point list was descriptive but lacked structure. This revision defines transparency operationally (what stakeholders can access), explains blockchain's specific technical role (hash commitments for integrity), and provides concrete implementation examples. The language shifts from "allows for understanding" to "ensures artifacts remain accessible," making the requirement testable.

Commented [19]: I'd also add Decentralized AI and Datacenter services to that list; see companies like www.arcblock.io and <https://www.xfra.ai/> for examples.



Examples:

- Model retraining event triggers blockchain transaction: timestamp, responsible actor, training data version hash, resulting model version hash
- Dataset update event triggers transaction: number of records changed, affected features, approver digital signature, reason for change
- Performance anomaly triggers alert: metric threshold exceeded, prediction accuracy drops, recommended corrective action, escalation timestamp

2.3. Traceability: Forensic Attribution and Chain-of-Custody

Definition: Each significant action in the AI system lifecycle can be traced to responsible parties, decisions can be audited back to training data and model versions, and root causes of failures can be investigated.

Blockchain Role: Digital signatures on every transaction establish attribution. Cryptographic linking (hash chains) establishes lineage between artifacts. Immutable records prevent tampering or erasure, enabling forensic investigation. Blockchain transactions recording model evolution **shall** include cryptographic hash commitments of the physical execution environment state. This includes verified hardware IDs, host driver variants, cluster orchestration parameters, and GPU/TPU library versions (e.g., CUDA/ROCm). True forensic reconstruction is impossible if the underlying hardware environment cannot be identically replicated.

Examples:

- Model output trace: Which model version generated this prediction? Which training data version was used? Which hyperparameters? All recorded as linked on-chain transactions enabling full reconstruction.
- Deployment decision audit: Who approved production deployment? What testing was conducted? What baseline metrics were required? Who is responsible if the system fails?
- Incident investigation: Trace failure back through model evolution changelog, training data version history, infrastructure changes, and approval signatures to identify root cause.

2.4. Security and Immutability

Definition: The integrity of AI system records, training data, and model artifacts is protected against unauthorized modification or deletion. Records serve as evidence for auditing and compliance verification.

Commented [20]: Original text mentioned "incentive-based attribution" rather than "forensic attribution" under traceability, without explaining what either one means or how it works. This revision clarifies traceability as chain-of-custody for forensic purposes (root cause analysis). The examples show how blockchain's cryptographic linking enables reconstructing the full history of a decision or failure, which is critical for accountability and continuous improvement. "Incentive-based attribution" is just as important and will be addressed elsewhere in the document (TBD).



Blockchain Role: Cryptographic hashing and distributed consensus make tampering computationally infeasible. Immutability prevents erasure or retroactive modification of audit records, ensuring evidence preservation.

Data-Compute Co-location Proofs: For Type 3 and high-consequence systems, blockchain roles extend to logging the physical proximity boundaries of data processing. Cryptographic validation tracks that sensitive or sovereign datasets remain processing-adjacent via ultra-high-speed bus interconnects and GPU-Direct Storage (GDS), preventing mid-transit data exposure outside authorized modular compute enclaves.

Important Caveat: Immutability **does not** mean "data cannot be changed." New entries are appended to the blockchain; historical records remain unchanged. This accommodates right-to-erasure requirements (see Section 4.5 on GDPR compliance) by recording erasure events and updating dataset version hashes rather than deleting original records.

2.5. Regulatory Compliance and Legal Frameworks

Definition: Blockchain provides cryptographic evidence and audit trails that demonstrate compliance with applicable regulations (GDPR, EU AI Act, sector-specific rules).

Blockchain Role: On-chain transaction logs serve as tamper-proof compliance documentation. Hash commitments to privacy-preserving artifacts (encrypted data, anonymized datasets) satisfy transparency requirements without exposing sensitive information.

Applicable Regulations:

- **GDPR (EU Regulation 2016/679):** Right to access, right to erasure, right to explanation, data minimization. (See Section 4.5 on hybrid on-chain/off-chain storage models.)
- **EU Artificial Intelligence Act (Regulation 2024/1689):** High-risk AI must maintain training data information, performance monitoring, human oversight records, incident reporting. (See Section 4 requirements aligned to AI Act Annex IV.)
- **NIST AI Risk Management Framework (AI RMF 1.0):** Organizations should document AI system provenance, design choices, testing procedures, deployment conditions, and post-deployment monitoring.
- **ISO/IEC 42001:2023:** AI Management Systems standard requiring governance, risk assessment, controls, and continuous improvement.
- **OECD AI Principles:** Transparency, accountability, robustness, fairness, human-centered values.

2.6. Standardization and Interoperability



Definition: Common standards for AI artifact formats, storage interfaces, and compliance documentation enable:

- Solutions from different providers to interoperate (e.g., models trained on one platform can be audited on another)
- Auditors and regulators to assess solutions consistently against common criteria
- Organizations to compare solutions against common benchmarks for procurement decisions

Blockchain Role: Standardized on-chain identifiers and metadata schemas enable cross-solution verification and integration. This supplement defines schemas for AI artifact registration (model hash, training data hash, performance metrics) that facilitate interoperability.

Examples:

- **Hardware and Infrastructure Provenance Schema:** [Node-ID][Accelerator-Type][Firmware-SHA256][Driver-Version][Orchestration-Manifest-Hash][Timestamp]
- *Rationale: Ensures that auditors can mathematically verify that the physical infrastructure layer meets the security baselines required for sovereign or critical infrastructure applications.*
- **Model versioning schema:**
[Model-ID][Hash-SHA256][Training-Data-Hash][Performance-Metrics][Approver-Signature]
- **Training audit trail schema:**
[Dataset-Version][Feature-Engineering-Method][Hyperparameters][Validation-Metrics][Timestamp]
- **Compliance declaration schema:**
[Regulation][Evidence-Location-URI][Verification-Status][Auditor-Signature][Timestamp]

3 AI Security Assessment Subject Matter Expert (AI SEC SME) Qualifications

This section describes the recommended training pathway for AI Security Assessment Subject Matter Experts (AI SEC SMEs), who are qualified to serve on BMM AI assessment teams. The training pathway ensures assessors have the necessary expertise to evaluate blockchain-enabled AI systems against the requirements in Section 4.



3.1. Overview

AI SEC SMEs must understand:

- AI system architecture, training methodologies, and common failure modes
- Attack vectors against AI systems (data poisoning, adversarial examples, model inversion, membership inference, prompt injection)
- Blockchain-based auditability mechanisms and their application to AI governance
- Applicable regulatory frameworks (GDPR, EU AI Act, NIST AI RMF, OECD AI Principles, ISO/IEC 42001)
- Assessment methodology: evaluating solutions against AI Supplement requirements using evidence-based approaches

3.2. Recommended Training Components

Module 1: AI Fundamentals (16 hours)

- **Machine Learning Lifecycle:** Data collection → preprocessing → training → validation → deployment → monitoring
- **Common Architectures:** Neural networks, decision trees, ensemble methods, large language models (LLMs), computer vision, natural language processing (NLP)
- **Black-box vs. Interpretable Models:** Explainability techniques (LIME, SHAP, attention mechanisms, saliency maps)
- **Performance Metrics:** Accuracy, precision, recall, F1, AUC-ROC, RMSE, bias/fairness metrics

Module 2: AI Security and Threat Models (20 hours)

- **Data Poisoning:** Label flipping, backdoor insertion, gradient-based attacks
- **Model Inversion & Membership Inference:** Extracting training data or determining data membership through adversarial queries
- **Adversarial Examples:** Evasion attacks (crafted inputs causing misclassification), transferability across models
- **Prompt Injection (LLMs):** Jailbreaking, prompt leaking, indirect prompt injection
- **Defense Mechanisms:** Input validation, anomaly detection, differential privacy, federated learning, adversarial training
- **Case Studies:** Real-world AI security incidents and lessons learned
- **Silicon and Firmware-Layer Threat Vectors:** Supply chain exploits targeting low-level components, including compromised Baseboard Management Controllers (BMCs), malicious Network Interface Card (NIC) firmware, and GPU/TPU firmware manipulation.



- **Hardware-Enforced Protections:** Hands-on evaluation of Confidential Computing implementations (e.g., AMD SEV, Intel TDX) to protect data-in-use within multi-tenant or non-trusted infrastructure environments.

Module 3: Blockchain and Auditability (12 hours)

- Distributed ledger fundamentals: consensus mechanisms, cryptographic hashing, digital signatures
- Smart contracts and transaction logs for AI lifecycle event recording
- On-chain vs. off-chain storage: trade-offs between immutability and privacy
- Privacy-preserving techniques: zero-knowledge proofs, homomorphic encryption, secure multi-party computation
- How blockchain enables tamper-proof audit trails and forensic analysis
- **Infrastructure Provenance Logging:** Frameworks for committing physical execution environment state schemas to the blockchain (recording verified Node-IDs, accelerator types, host driver variants, and library versions like CUDA/ROCm) to enable true deterministic replay.
- **Data-Compute Proximity Auditability:** Auditing the physical co-location boundaries of sensitive data via ultra-high-speed bus interconnects and GPU-Direct Storage (GDS) to cryptographically verify data sovereignty and eliminate mid-transit intercept risks.

Commented [21]: Cross-check with existing BMM assessment training to see if it is even needed.

Module 4: AI Regulation and Governance (16 hours)

- **GDPR:** Data minimization, right to erasure, transparency, data protection impact assessments (DPIAs)
- **EU Artificial Intelligence Act:** Risk classification, conformity assessment, high-risk requirements (Annex IV documentation), post-market surveillance
- **NIST AI RMF:** Map, Measure, Manage, Govern functions; AI lifecycle risk management
- **ISO/IEC 42001:2023:** AI Management Systems—governance structure, risk assessment, controls, improvement cycles
- **OECD AI Principles:** Transparency, accountability, robustness, safety, fairness

Module 5: Assessment Methodology (20 hours)

- GBA BMM assessment framework and procedures (adapted for AI)
- Evaluating solutions against Section 4 requirements: evidence collection, documentation review, interviews, demonstrations
- Risk-type determination (Type 1, 2, 3) and application to requirement interpretation
- Consensus-building among assessment team members; handling disagreements
- Domain context and feasibility considerations (not all requirements apply uniformly)



Module 6: Hands-On Assessment Workshop (16 hours)

- Supervised assessment of a real or realistic blockchain-enabled AI system
- Practice evidence evaluation, artifact verification (hash checking, signature validation), team consensus
- Mock interviews with solution developers and compliance officers
- Draft findings, identify gaps, propose remediation recommendations
- Debriefing with lead assessor: strengths, areas for improvement, calibration
- **Decentralized Enclave & Boundary Auditing:** Evaluation of modular, containerized micro-datacenter units (field-deployable AI pods) operating outside traditional hyperscalers, testing for physical perimeter security constraints, thermal/power resilience protocols, and automated secure hardware-erase capabilities.

Total Training Duration: ~100 hours (equivalent to 2.5 weeks full-time or 8–10 weeks part-time)

3.3. Qualification Requirements

To be qualified to serve as an **AI SEC SME** on a BMM AI assessment team, an individual must:

- 1.** Complete the full training pathway (100 hours minimum) through a GBA-accredited training provider OR demonstrate equivalent experience (see waiver provision below)
- 2.** Pass the AI SEC Knowledge Assessment (written exam covering all six modules, minimum 75% passing score, proctored)
- 3.** Demonstrate domain knowledge in at least one of: AI system development, AI security research, blockchain/distributed ledger technology, or regulatory compliance
- 4.** Participate in at least one supervised assessment (apprenticeship model under a certified Lead Assessor)

Maintain active GBA BMM professional membership and complete annual continuing education (minimum 8 hours/year on AI governance topics)

Waiver Provision:

Individuals with advanced degrees (MS, PhD) in Computer Science, AI, Data Science, or related fields, plus 3+ years of professional AI experience, may request a waiver of the training hours requirement. The waiver requires:

- Submission of CV demonstrating relevant education and experience
- Documented demonstration of knowledge in all six modules (portfolio, publications, project work)
- Passing score (85%+) on the AI SEC Knowledge Assessment
- Lead assessor approval and GBA Director of Standards & Certification endorsement



Purpose: This waiver pathway ensures the program scales efficiently by recognizing existing AI expertise, avoiding bottlenecks on training capacity.

4 Solution Requirements

This section defines the requirements that a blockchain-enhanced AI system must address to be considered for BMM AI certification. As noted in the Introduction, this supplement serves two distinct audiences — **solution implementers** and **solution assessors** — and the requirements in this section are designed to speak to both, though in different ways.

- **For Solution Implementers**

Requirements are expressed using RFC 2119 terminology:

- **SHALL** — The solution must satisfy this requirement. Absence of compliance is a disqualifying deficiency.
- **SHOULD** — The solution is strongly recommended to satisfy this requirement. Non-compliance is permitted but must be accompanied by a documented rationale explaining why the requirement was not adopted and what mitigating measure, if any, was put in place.
- **MAY** — The solution may optionally satisfy this requirement. It represents a recognized best practice or forward-looking capability that is neither expected nor penalized if absent.

Implementers should treat **SHALL**, **SHOULD**, and **MAY** as a complete picture of what a well-governed, trustworthy AI system looks like — from the baseline minimum to the aspirational ideal. Building toward **SHOULD** and **MAY** requirements, even where not strictly required for certification, strengthens the solution's overall governance posture and reduces the likelihood of gaps emerging during assessment or post-deployment.

- **For Solution Assessors**

Assessors focus their compliance determination on **SHALL** requirements only.

A **SHOULD** or **MAY** requirement is not a basis for a non-compliant finding; however, assessors may note the absence of **SHOULD** requirements as an observation or improvement opportunity in the assessment report without it affecting the certification determination.

For each applicable **SHALL** requirement, the assessor determines one of four outcomes:

- **Compliant** — Evidence clearly demonstrates the requirement is fully satisfied.
- **Partially Compliant** — The requirement is substantially addressed but a specific, bounded gap exists. A remediation plan is required.



- **Non-Compliant** — The requirement is unaddressed or the gap is material and undermines the control's effectiveness. Certification cannot be awarded at the stated level until remediated.
- **Not Applicable** — The requirement does not apply to this solution's risk type or domain context, with documented justification.

The risk type of the solution (Type 1, 2, or 3, as defined in Section 1.2) governs which requirements carry a **SHALL** designation for that solution. Assessors should establish the risk type before beginning requirement-by-requirement review, as this directly determines the scope of mandatory compliance.

Applicability Key

Symbol	Meaning
SHALL	Mandatory for certification for this type
SHOULD	Strongly recommended; not mandatory
MAY	Optional / forward-looking
—	Not applicable to this type

4.1 Requirement Language

ReqID	Description	Type 1	Type 2	Type 3	Comments
4.1.1	All requirements use RFC 2119 language: "SHALL" = mandatory for certification; "SHOULD" = strongly recommended; "MAY" = optional	SHALL	SHALL	SHALL	Uniform across all types. Establishes interpretation rules for all subsequent requirements; eliminates ambiguity in assessment determinations.

4.1.2 Architecture-Neutral Application of Requirements (mandatory interpretive rule)

The requirements in Section 4 express underlying assurance outcomes: reproducible evaluation, traceable provenance, tamper-evident accountability, controlled artifact access, and statistically valid performance evidence. Specific techniques named in individual requirements — including

Commented [22]: Section 4.2.1 makes open-source publication a SHALL requirement for Type 1 systems, while Section 4.4.1 describes open-source licensing as SHOULD.

These two provisions appear inconsistent.

A low-risk system may still contain legitimate proprietary intellectual property. Risk level alone should not determine whether source code must be publicly released.

I recommend consolidating Sections 4.2 and 4.4 into one disclosure model:

Option A: Public source code;
Option B: Restricted auditor access;
Option C: Exceptional hash-only disclosure for justified critical or classified systems.

I need some time today to edit the whole section and will make it in separate doc

Commented [23]: Thanks for catching that. A legitimate inconsistency. I will merge the two tables in some fashion.



train/validation/test splits, cross-validation, model-weight hashing, and blockchain transaction recording — are acceptable methods of satisfying the underlying assurance outcome for conventional trained-model architectures but are not the exclusive means of compliance.

For solutions built on architectures where a named technique does not natively apply — including third-party foundation-model APIs, retrieval-augmented generation (RAG) systems, agentic systems, continuously learning or online-learning systems, federated-learning systems, and closed third-party models — the solution provider shall propose, and the assessment team shall evaluate, an equivalent evidence method that satisfies the same assurance outcome. Examples include held-out evaluation sets and documented prompt/response logs in place of train/test splits, API-level usage and output logging in place of model-weight hashing, and retrieval-corpus versioning with hash commitments in place of training-data hashing. The assessment report shall document any substituted method and the rationale for its equivalence. This clause does not waive any SHALL requirement; it governs how the requirement's assurance outcome may be evidenced.

4.2 Source Code Availability and Transparency

Applicability note: Requirements 4.2.1 through 4.2.3 describe alternative source-code disclosure pathways. A solution need only satisfy the pathway or combination of pathways applicable to its risk type and circumstances. Requirements 4.2.4 through 4.2.9 apply independently.

ReqID	Description	Type 1	Type 2	Type 3	Comments
4.2.1	Option A — Public source-code availability: Source code is published in a version-controlled repository under an OSI-approved or AI-specific license. The publication includes the version history necessary for independent review. License selection is governed by Section 4.4.	SHOULD	SHOULD	MAY	Public availability provides the highest transparency and facilitates independent review. It is strongly recommended for Type 1 and Type 2 systems where intellectual-property, safety, legal, or national-security constraints do not prohibit it. For Type 3 systems, public disclosure may create an



					unacceptable attack surface or be prohibited; it is therefore optional.
4.2.2	Option B — Restricted source-code availability: Source code is made available to authorized auditors, regulators, and, where appropriate, customers under an NDA, security clearance, audit agreement, or equivalent controlled-access arrangement. The solution documents the access process, eligibility criteria, contact method, and typical response time. Legitimate trade secrets may be redacted only where the redaction does not prevent meaningful assessment.	SHOULD	SHALL	SHALL	This is the minimum acceptable disclosure pathway for Type 2 systems and the default pathway for Type 3 systems that cannot safely publish source code. For Type 1 systems, it is an acceptable alternative to Option A. Type 3 systems may apply further access controls, but must preserve meaningful independent assessment.
4.2.3	Option C — Exceptional hash-only disclosure: Source code is not made available publicly or through ordinary controlled access. A cryptographic hash, timestamp, and public-key	—	—	MAY	Permitted only for Type 3 systems where disclosure would conflict with documented national-security classification, legal restriction, critical-safety protection, or similarly



	identifier are recorded in the tamper-evident lifecycle record. During assessment, the authorized assessor verifies the hash against the source code under a controlled environment or equivalent approved procedure.				compelling justification. The exception requires Lead Assessor approval and GBA Director of Standards & Certification approval. Hash-only disclosure does not eliminate the requirement for meaningful independent assurance.
4.2.4	Source-code provenance record maintained: Each released source-code version has a unique identifier, cryptographic hash, timestamp, and authorized approver signature recorded in the solution's tamper-evident lifecycle record.	SHALL	SHALL	SHALL	Applies regardless of disclosure pathway. This establishes integrity, version traceability, and evidence that the assessed code corresponds to the deployed code. See the common artifact provenance control in Section 4.6.20.
4.2.5	Architecture and design documentation provided, including system components, data-flow diagrams, interfaces, trust boundaries, and deployment architecture.	SHALL	SHALL	SHALL	Documentation is required even where source code is unavailable or restricted. It gives assessors enough context to evaluate design choices, data flows, dependencies, and security boundaries.
4.2.6	Deployment, configuration, operation, and rollback instructions provided at a	SHALL	SHALL	SHALL	The documentation may be access-controlled for Type 3 systems, but it must remain



	level sufficient for authorized parties to deploy, operate, maintain, and assess the solution safely.				available to authorized assessors.
4.2.7	Known limitations, operating constraints, unsupported scenarios, material dependencies, and known defects documented.	SHALL	SHALL	SHALL	This includes material performance boundaries, security limitations, failure conditions, and situations in which the system should not be used.
4.2.8	Software Bill of Materials (SBOM) provided in SPDX, CycloneDX, or another recognized machine-readable format. The SBOM identifies material dependencies, versions, suppliers where known, and applicable licenses.	SHOULD	SHALL	SHALL	For Type 2 and Type 3 systems, the SBOM is required to support supply-chain risk assessment, vulnerability management, provenance verification, and license compliance.
4.2.9	Source-code changes included in the common lifecycle change-control record required by Section 4.3.11. Each applicable entry identifies the version, material change, reason for change, date and timestamp, and approving authority.	SHALL	SHALL	SHALL	This requirement covers source-code-specific evidence within the common lifecycle change-control process. It does not create a separate duplicate change-log process.

4.3 AI **Model** Availability and Versioning

ReqID	Description	Type 1	Type 2	Type 3	Comments
4.3.1	Model type documented (e.g., neural network, decision tree, random forest, LLM, computer vision)	SHALL	SHALL	SHALL	Uniform across all types. Baseline metadata; required for any meaningful audit regardless of risk level.
4.3.2	Model architecture described: layers, activation functions, parameter count, complexity metrics, training framework and version (e.g., PyTorch 2.0, TensorFlow 2.15)	SHALL	SHALL	SHALL	Uniform across all types. Enables assessors to evaluate appropriateness of model for stated use case at every risk level.
4.3.3	Summary statistics documented: model file size, inference latency (in ms), memory requirements (RAM/GPU VRAM)	SHALL	SHALL	SHALL	Uniform across all types. Enables verification of deployment feasibility and performance claims regardless of risk.
4.3.4	Intended use documented: approved use cases and explicitly prohibited uses	SHALL	SHALL	SHALL	Uniform across all types. Scope limits prevent misuse and provide accountability anchor at every risk level.
4.3.5	Known limitations documented: performance boundaries, out-of-	SHALL	SHALL	SHALL	Uniform across all types. Critical for responsible deployment; maps to EU AI

Commented [24]: "Model": BMM model or something else? Be specific.

There are several types of AI Models: e.g., Foundation, Frontier, General Purpose. Which one?

Commented [25]: AI Model used in the system under assessment. Qualified model in the section heading.



	distribution scenarios where model is unreliable, known failure modes				Act Annex IV requirements for every type.
4.3.6	Each new model version assigned unique identifier (UUID or semantic versioning e.g., "v1.2.3")	SHALL	SHALL	SHALL	Uniform across all types. Foundational for version traceability across all downstream requirements.
4.3.7	Model provenance record maintained: each released model binary shall have a model identifier, cryptographic hash, training-data version reference, performance summary, timestamp, and authorized approver signature recorded in the solution's tamper-evident lifecycle record.	SHALL	SHALL	SHALL	Artifact-specific fields are governed by the common Artifact Provenance and Lifecycle Control (4.6.20). This row identifies the model artifact to which that control applies.
4.3.8	Model artifact made available via Option A : Binary available via decentralized/public storage (IPFS, Hugging Face, public download). Stakeholders can verify integrity by comparing downloaded hash to on-chain record.	SHOULD	SHOULD	MAY	Type 1–2: Preferred for maximum transparency and to enable external red-teaming and research validation, but Option B is acceptable. Type 3: Public availability of model weights for critical infrastructure systems may create unacceptable attack surface (adversaries can study the model offline); hash-only or



					restricted access is more appropriate.
4.3.9	Model artifact made available via Option B : Binary restricted to authorized parties (auditors, regulators, customers under NDA). Hash commitment published on-chain. Access request process documented.	MAY	SHALL	SHOULD	Type 1 : Optional; open publication (Option A) is preferred for low-risk systems. Type 2 : Mandatory as minimum acceptable transparency; full public availability is preferred but conditional auditor access is sufficient for certification. Type 3 : Recommended alongside the hash-only option (Option C); auditor access should remain available even when public disclosure is restricted.
4.3.10	Model artifact made available via Option C : Binary not disclosed. Hash commitment on-chain only. Auditor verifies integrity during off-chain assessment under strict NDA.	—	—	MAY	Type 1–2 : Not acceptable. Auditor access to model weights is required for meaningful assessment of high-risk systems. Type 3 : Permissible for critical infrastructure where model disclosure could enable targeted adversarial attacks or violates national security constraints; must be explicitly



					justified in the assessment report.
4.3.11	Common lifecycle change-control record maintained for source code, datasets, models, model weights, and deployment configurations. Each relevant entry shall identify the artifact and version; date and timestamp; what changed; reason; performance or operational impact where applicable; and approver authorization. Material entries shall be recorded in the tamper-evident lifecycle record.	SHOULD	SHALL	SHALL	Consolidates the common change-control elements formerly repeated in 4.2.9 and 4.5.8. Artifact-specific entries remain required through the cross-references in those rows.

4.4 Open-Source Licensing, Interoperability, and Process Transparency

Applicability note: Section 4.4 does not independently require public source-code disclosure. It governs the licensing, interoperability, and public process-transparency obligations associated with the disclosure pathway selected under Section 4.2.

ReqID	Description	Type 1	Type 2	Type 3	Comments
4.4.1	Where the solution elects Option A under Section 4.2.1 , source code shall be released under a documented OSI-approved or AI-specific license appropriate to the use case. Examples include Apache 2.0, MIT, BSD 3-Clause, MPL 2.0,	SHOULD	SHOULD	MAY	This requirement governs the quality and clarity of the license if public release is chosen . It does not require a Type 1 or Type 2 system to select Option A. The selected license should clearly state permissions,

Commented [26]: "Open Source" missing a noun. AI or blockchain?

See MOF, (AI Model Openness Framework), a comprehensive system for evaluating and classifying completeness and openness of AI ML models. MOF assesses which components of AI model development lifecycle are publicly released and under what licenses, ensuring an objective evaluation.

<https://isitopen.ai/>



	GPL 3.0, AGPL 3.0, OpenRAIL, BigCode Open RAIL, or BLOOM RAIL.				restrictions, attribution obligations, derivative-work treatment, and any responsible-use limitations.
4.4.2	Where the solution elects Option B under Section 4.2.2 , the solution shall document the controlled-access terms and interoperability information necessary for meaningful assessment and authorized integration. This includes, as applicable, APIs, data schemas, interface specifications, integration guides, and a justification for not using public release.	SHOULD	SHALL	SHALL	For Type 2 and Type 3 systems using restricted disclosure, this control ensures that proprietary treatment does not result in opaque or non-interoperable deployment. The justification should identify the applicable IP, safety, security, legal, or national-security basis.
4.4.3	Where the solution elects Option C under Section 4.2.3 , alternative transparency and assurance mechanisms shall be documented. These mechanisms may include a controlled assessor review environment, independent third-party assessment, reproducible build evidence, artifact provenance records, interface documentation, or other measures sufficient to support meaningful assurance.	—	—	MAY	This requirement applies only to exceptional Type 3 hash-only disclosure cases. It is not an alternative pathway for Type 1 or Type 2 systems. Approval requirements remain governed by Section 4.2.3.
4.4.4	Risk-management and governance protocols publicly documented at a level that does not expose protected implementation details. At	SHALL	SHALL	SHALL	Process transparency is distinct from source-code disclosure. A solution may protect proprietary or classified code while still



	minimum, documentation shall address data governance, validation and testing approach, monitoring and alert thresholds, human oversight where applicable, incident response, audit-trail retention, and the process for reporting material issues.				providing meaningful information about how the system is governed, tested, monitored, and held accountable.
4.4.5	License status, disclosure pathway, and applicable access restrictions recorded as metadata associated with each versioned source-code and model artifact. Material license or access-condition changes shall be recorded in the common lifecycle change-control record.	SHALL	SHALL	SHALL	This creates auditable evidence of whether an artifact is publicly released, restricted, or subject to an approved exceptional disclosure approach. It also prevents undocumented retroactive changes to license or access terms.
4.4.6	The solution documents interoperability dependencies and constraints, including supported interfaces, data formats, protocols, version-compatibility requirements, and known integration limitations relevant to authorized deployment, audit, or migration.	SHOULD	SHALL	SHALL	This preserves the “interoperability” function of Section 4.4. Type 2 and Type 3 systems must enable assessors and authorized integrators to understand material dependencies and integration risks.

4.5 Dataset Availability and GDPR Compliance

ReqID	Description	Type 1	Type 2	Type 3	Comments
4.5.A – Dataset Metadata					
4.5.1	Data sources documented: how collected (public	SHALL	SHALL	SHALL	Uniform across all types. Enables bias and



ReqID	Description	Type 1	Type 2	Type 3	Comments
	datasets, proprietary databases, user submissions, web scraping, synthetic data, sensors), with provenance for each source				representativeness assessment at every risk level. Maps to EU AI Act Annex IV §2(c).
4.5.2	Dataset volume and distribution documented: number of records, features, data types, class distribution (balanced or imbalanced)	SHALL	SHALL	SHALL	Uniform across all types. Required to assess adequacy of training data for stated use case regardless of risk level.
4.5.3	Temporal coverage documented: date range of data collection, time period characteristics (time series, snapshot, continuous)	SHALL	SHALL	SHALL	Uniform across all types. Distribution shift risk increases when deployment period diverges from training period; this is a universal concern.
4.5.4	Geographic and demographic coverage documented: which populations are represented, which are underrepresented or excluded	SHOULD	SHALL	SHALL	Type 1: Recommended as fairness best practice but demographic gaps in non-consequential systems carry lower harm potential. Type 2–3: Mandatory because underrepresented groups face elevated model error rates in high-stakes decisions (lending, healthcare, hiring), creating discriminatory outcomes that must be assessable.
4.5.5	Data quality metrics documented: missing value	SHOULD	SHALL	SHALL	Type 1: Good practice but data quality



ReqID	Description	Type 1	Type 2	Type 3	Comments
	rates, outlier statistics, consistency checks, duplicate detection results				deficiencies in low-risk systems have limited consequences. Type 2–3: Mandatory because poor data quality directly degrades the reliability of decisions affecting individual rights or safety; auditors must be able to assess whether training data met acceptable quality thresholds.
4.5.6	Labeling methodology documented (if supervised learning): who labeled, platform, inter-annotator agreement score (Cohen's kappa or equivalent), label guidelines and edge case handling	SHOULD	SHALL	SHALL	Type 1: Recommended for completeness but label quality issues in low-risk systems have bounded impact. Type 2–3: Mandatory because label quality directly determines model reliability; high-risk decisions require demonstrably high label quality, and inter-annotator agreement provides objective evidence of label consistency.
4.5.7	Dataset provenance record maintained: each dataset version shall have a unique identifier, cryptographic hash, timestamp, and authorized approver signature recorded in the	SHALL	SHALL	SHALL	Artifact-specific fields are governed by the common Artifact Provenance and Lifecycle Control (4.6.20). This row identifies the dataset



ReqID	Description	Type 1	Type 2	Type 3	Comments
	solution's tamper-evident lifecycle record.				artifact to which that control applies.
4.5.8	Dataset changes shall be included in the common lifecycle change-control record required by 4.3.11, including records added, removed, or corrected; reason; timestamp; and authorizer signature.	SHALL	SHALL	SHALL	Consolidated with 4.3.11 to eliminate duplicate lifecycle change-log requirements while preserving dataset-specific evidence.
4.5.B – GDPR Compliance					
4.5.9	PII stored off-chain only (encrypted database, secure cloud storage, data clean room, or access-controlled secure enclave). Only metadata and cryptographic hash of dataset stored on-chain.	SHALL	SHALL	SHALL	Uniform across all types. Resolves the blockchain immutability vs. GDPR conflict at every risk level. On-chain stores the audit proof; off-chain stores the mutable data subject to erasure rights.
4.5.10	Hybrid on-chain/off-chain storage implemented (preferred GDPR approach): On-chain stores hash + metadata + approver signature. Off-chain stores actual dataset. Upon right-to-erasure request: delete PII from off-chain, compute new dataset hash, record version update transaction on-chain with reason, privacy officer signature, and timestamp.	SHOULD	SHALL	SHALL	Type 1: Strongly recommended because GDPR applies regardless of risk type, but enforcement scrutiny on low-risk systems is lower. Type 2–3: Mandatory because high-risk systems processing personal data will face GDPR audits; documented erasure procedures are essential for regulatory compliance.
4.5.11	Cryptographic erasure as alternative GDPR approach:	MAY	MAY	MAY	Optional across all types. Acceptable where



ReqID	Description	Type 1	Type 2	Type 3	Comments
	sensitive records encrypted with per-record AES-256 keys; upon erasure request, encryption key deleted rendering data permanently inaccessible. On-chain transaction records: key deletion event, timestamp, authorizer signature, affected record count.				physical deletion is impractical (distributed backups, append-only storage). Legal counsel should confirm acceptability in applicable jurisdictions before relying on this approach; some data protection authorities accept cryptographic erasure; others require physical deletion.
4.5.12	Auditor access mechanism documented for datasets containing PII: contact method, eligibility criteria, NDA/data-sharing agreement requirements, typical response time	SHALL	SHALL	SHALL	Uniform across all types. Enables compliance verification without public data exposure at every risk level.
4.5.13	Data Protection Impact Assessment (DPIA) conducted and available to auditors where dataset contains sensitive personal data (health, financial, biometric, criminal justice)	—	SHALL	SHALL	Type 1: Not required; low-risk systems are unlikely to meet the GDPR Art. 35 threshold for mandatory DPIA. Type 2–3: Mandatory because processing sensitive personal data for high-stakes decisions is precisely the high-risk processing scenario that GDPR Art. 35 targets; DPIA is a legal requirement, not merely best practice.



ReqID	Description	Type 1	Type 2	Type 3	Comments
4.5.14	Data-to-Compute Co-location: Architectures handling highly sensitive, sovereign, or classified datasets <i>shall</i> utilize high-speed bus interconnects and GPU-Direct Storage (GDS) to process data locally within verified boundaries.		SHOULD	SHALL	Type 1–2: Discretionary based on data sensitivity. Type 3: Mandatory to eliminate mid-transit intercept and data exposure risks outside authorized modular compute enclaves.

4.6 Training Loop Availability



ReqID	Description	Type 1	Type 2	Type 3	Comments
4.6.A – Training Methodology					
4.6.1	Training algorithm(s) documented (e.g., SGD, Adam, genetic algorithm, reinforcement learning variant)	SHOULD	SHALL	SHALL	Type 1: Recommended for completeness but algorithm transparency has limited audit value for non-consequential decisions. Type 2–3: Mandatory because algorithm choice affects model properties (convergence behavior, susceptibility to adversarial inputs, bias amplification) that are material to high-risk assessments.
4.6.2	Loss function(s) and optimization objectives documented (cross-entropy, MSE, custom; single- or multi-objective)	SHOULD	SHALL	SHALL	Type 1: Recommended as good documentation practice. Type 2–3: Mandatory because the loss function reveals what the model was actually optimized for, which may differ from the stated purpose — a critical accountability check for high-risk systems.



4.6.3	Training duration and computational resources documented (epochs, wall-clock time, hardware used, e.g., "50 epochs, 48 hours, 8× NVIDIA A100, ~3,200 GPU-hours")	SHOULD	SHOULD	SHALL	Type 1–2: Useful context but not critical for assessment conclusions. Type 3: Mandatory because reproducibility verification for critical systems requires knowing the exact computational environment; hardware differences can cause non-deterministic outcomes that must be accounted for.
4.6.4	Weight-initialization strategy documented. Where reproducibility is required, associated random seed shall be included in the reproducibility and execution-environment documentation required by 4.6.18.	SHOULD	SHALL	SHALL	Cross-reference added to avoid duplicate seed documentation requirements.
4.6.B – Hyperparameters					



4.6.5	Learning rate and schedule documented (initial value, decay strategy: constant, step, cosine annealing)	SHOULD	SHALL	SHALL	Type 1: Recommended but not assessed as a compliance gate for low-risk systems. Type 2–3: Mandatory because learning rate is the most impactful hyperparameter; its documentation is required for meaningful reproduction and for assessing whether training was conducted responsibly.
4.6.6	Batch size and gradient accumulation steps documented	SHOULD	SHALL	SHALL	Type 1: Useful documentation but not material to low-risk assessments. Type 2–3: Mandatory because batch size affects training stability, model generalization, and memory requirements; auditors must verify training configuration was appropriate for the data volume and model complexity.



4.6.7	Regularization parameters documented (L1/L2 strength, dropout rate, weight decay)	SHOULD	SHALL	SHALL	Type 1: Recommended but low consequence if undocumented for non-critical systems. Type 2–3: Mandatory because regularization choices directly affect overfitting risk and out-of-distribution generalization — both critical concerns for high-stakes deployments.
4.6.8	Data augmentation techniques documented if applied (random crops, rotations, mix-up, cutout, etc.)	SHOULD	SHOULD	SHALL	Type 1–2: Recommended but augmentation documentation is not a certification gate for these assessments. Type 3: Mandatory because augmentation significantly affects robustness and out-of-distribution behavior; for safety-critical systems these properties must be fully traceable and verifiable.



4.6.9	Stopping criteria documented (early stopping threshold, maximum epochs, convergence criteria)	SHOULD	SHALL	SHALL	Type 1: Good documentation practice but not a compliance gate. Type 2–3: Mandatory because stopping criteria determine whether a model was adequately trained; undocumented stopping means auditors cannot verify that training was completed appropriately or prematurely halted.
4.6.10	For LLMs specifically: context length, vocabulary size, tokenization method (BPE, WordPiece, SentencePiece) documented	SHOULD	SHALL	SHALL	Type 1: Relevant documentation but limited audit consequence for non-critical LLM applications. Type 2–3: Mandatory because these parameters define the capability boundaries and prompt injection risk surface of LLMs used in high-risk or safety-critical contexts.
4.6.C – Data Splits and Validation					



4.6.11	Train/validation/test split ratios documented with stratification method if applicable	SHALL	SHALL	SHALL	Uniform across all types. Required to assess validity of reported performance metrics regardless of risk level; without split documentation, performance claims cannot be independently evaluated.
4.6.12	Cross-validation strategy documented where applicable (k-fold, time-series split, leave-one-out)	SHOULD	SHALL	SHALL	Type 1: Recommended for thorough documentation but not a certification requirement. Type 2–3: Mandatory because cross-validation methodology affects the reliability of performance estimates; assessors must verify that validation was robust, especially for small or imbalanced datasets common in specialized domains.



4.6.13	Data-shuffling and stochastic-preprocessing settings documented. Where reproducibility is required, associated random seeds shall be included in the reproducibility and execution-environment documentation required by 4.6.18.	SHOULD	SHALL	SHALL	Cross-reference added to avoid duplicate seed documentation requirements.
4.6.D – Evaluation Metrics					
4.6.14	Primary performance metrics reported (accuracy, precision, recall, F1, AUC-ROC, RMSE, perplexity, etc. as appropriate to task)	SHALL	SHALL	SHALL	Uniform across all types. Minimum transparency requirement; must be sufficient for independent performance assessment regardless of risk level.



4.6.15	Secondary/fairness metrics reported (demographic parity, equalized odds, calibration by group)	SHOULD	SHALL	SHALL	Type 1: Recommended as fairness best practice but fairness gaps in non-consequential systems carry limited harm potential. Type 2–3: Mandatory because differential performance across protected groups in high-risk decisions (lending, hiring, healthcare) constitutes discrimination; these metrics must be documented, assessed, and defensible to regulators.
4.6.16	Baseline comparison provided (human performance, prior model version, or industry benchmark)	SHOULD	SHALL	SHALL	Type 1: Good context for performance claims but not a certification gate. Type 2–3: Mandatory to prevent inflated performance claims; auditors must be able to assess whether reported improvements are genuine and whether the model meaningfully outperforms existing approaches or human judgment.



4.6.17	Performance thresholds for production deployment defined and documented (e.g., "Minimum 92% accuracy; max 5% FPR; demographic parity difference <0.05")	SHALL	SHALL	SHALL	Uniform across all types. Without defined thresholds, deployment decisions are ungovernable at every risk level; thresholds are the mechanism by which governance intentions become enforceable gates.
4.6.E – Reproducibility and Audit Trail					
4.6.18	Reproducibility and execution-environment documentation maintained with sufficient detail for independent replication within expected stochastic variance. The documentation shall include applicable code version, dataset version, material hyperparameters, random seeds, known sources of non-determinism, and execution-environment provenance (hardware/accelerator type, relevant driver and library versions, and orchestration configuration).	—	SHOULD	SHALL	Consolidates the common reproducibility evidence previously distributed across 4.6.4, 4.6.13, and 4.6.21. Type 2 and Type 3 applicability is unchanged.



4.6.19	Git commit hash (or equivalent) of training codebase recorded on-chain for each training run	SHOULD	SHALL	SHALL	Type 1: Recommended as good version control hygiene. Type 2–3: Mandatory because the commit hash links the model version to the exact code state used to produce it; without it, full lifecycle reconstruction is impossible and forensic root-cause analysis is severely limited.
4.6.20	Each training run recorded as blockchain transaction: start/end timestamps, dataset version hash, model version hash, key hyperparameters subset, final validation metrics, authorizer signature, deviations from standard procedure	SHALL	SHALL	SHALL	Uniform across all types. Immutable training audit trail is a core requirement at every risk level; enables forensic root-cause analysis (e.g., "Why did model v3.2 underperform? Training record shows it used stale Dataset-v2.0 missing Q1 2026 data.")



4.6.21	Execution-environment provenance shall be documented as part of the reproducibility and execution-environment documentation required by 4.6.18. Evidence shall include, as applicable, Node-IDs or equivalent, accelerator types, driver variants, library versions, and orchestration configuration.		SHOULD	SHALL	Consolidated with 4.6.18 while retaining the infrastructure-provenance evidence expected for Type 2 and Type 3 systems.
4.6.22	Distributed Training State Auditing: Solutions utilizing multi-node distributed training <i>shall</i> immutably log cross-node synchronization parameters, distributed checkpoints, and cluster orchestration failover boundaries.		MAY	SHALL	Type 1–2: Optional. Type 3: Crucial for safety-critical systems to verify that node failures or synchronization lag during large-scale training runs did not introduce silent corruption or non-deterministic behavior.

4.7 AI Model Weights and Artifacts

ReqID	Description	Type 1	Type 2	Type 3	Comments
4.7.1	Model-weight provenance record maintained: each released model-weights artifact shall have a	SHALL	SHALL	SHALL	Artifact-specific fields are governed by the common Artifact Provenance and Lifecycle Control (4.6.20). This row identifies the

Commented [27]: BMM Model or AI?

Commented [28]: Model refers to AI Model. Not aware of any BMM model. Added AI.



ReqID	Description	Type 1	Type 2	Type 3	Comments
	cryptographic hash, timestamp, and authorized approver signature recorded in the solution's tamper-evident lifecycle record.				model-weights artifact to which that control applies.
4.7.2	Model binary stored in immutable or write-protected storage (append-only, write-once media, or versioned repository with modification access controls)	SHALL	SHALL	SHALL	Uniform across all types. Prevents retroactive tampering with deployed model weights; required at every risk level to ensure the on-chain hash continues to match the actual deployed model.
4.7.3	Model weights made available to authorized auditors on request; access request process documented (contact method, NDA/agreement requirements, ≤5 business day response for Type 3)	SHOULD	SHALL	SHALL	Type 1: Recommended to enable spot-check verification but not a certification gate for non-critical systems. Type 2–3: Mandatory because auditors must be able to independently verify the hash by computing it from the received weights; without weight access, model integrity cannot be confirmed and the audit is incomplete. Response time is more tightly constrained at Type 3 (≤5 business days) given the severity of potential compromise.
4.7.4	Chain of custody log maintained for model weights access: who accessed, when, for what purpose (view-only, exported, used in testing). Log tamper-	—	SHOULD	SHALL	Type 1: Not required; chain of custody overhead is disproportionate for low-risk applications. Type 2: Strongly recommended but not a certification gate; provides valuable accountability evidence. Type 3: Mandatory because



ReqID	Description	Type 1	Type 2	Type 3	Comments
	proof (on-chain or append-only system).				critical infrastructure systems must demonstrate that no unauthorized access or modification occurred between training and deployment; an unverified chain of custody creates unacceptable risk of weight substitution attacks.
4.7.5	Each model version retains full version history; old versions preserved and accessible (not overwritten)	SHALL	SHALL	SHALL	Uniform across all types. Enables rollback to a previous version upon performance degradation or security incident regardless of risk level.
4.7.6	Hardware Bill of Materials (HBOM): The solution <i>shall</i> maintain a documented provenance log of AI accelerators, including manufacturer cryptographic signatures, secure boot configuration status, and physical chain of custody.		SHOULD	SHALL	Type 1: Not required. Type 2–3: Necessary to mitigate physical and silicon-layer supply chain exploits, ensuring the underlying compute units have not been intercepted or modified.
4.7.7	Firmware Integrity Validation: Continuous runtime validation and secure-flash protections <i>shall</i> be enforced for critical low-level components (BIOS, BMC, NIC, and GPU firmware).		SHOULD	SHALL	Type 1: Not required. Type 2–3: Protects against persistent, low-level hardware exploits that evade standard operating system and software-level endpoint detection.
4.7.8	Decentralized Enclave & Boundary Auditing: For deployments operating outside centralized cloud	MAY	SHOULD	SHALL	Type 1–2: Discretionary. Type 3: Mandatory for tactical edge, air-gapped, or critical infrastructure systems where physical



ReqID	Description	Type 1	Type 2	Type 3	Comments
	hyperscalers, the infrastructure <i>shall</i> satisfy verified standards for perimeter physical security, thermal/power resilience, and automated secure-erase capabilities (e.g., modular, field-deployable AI pods).				tampering, environmental failure, or disconnected states pose operational risks.

4.8 All **Actors'** Signatures and Accountability

ReqID	Description	Type 1	Type 2	Type 3	Comments
4.8.A – What Must Be Signed					
4.8.1	Model training approval digitally signed by authorized approver before training begins	SHALL	SHALL	SHALL	Uniform across all types. Establishes accountability for the decision to train regardless of risk level; prevents unauthorized or ad hoc training runs.
4.8.2	Model deployment to production digitally signed; multi-signature quorum required (e.g., ML Lead + Security + Legal = 3-of-3). Quorum configuration documented in governance policy.	SHALL	SHALL	SHALL	Uniform across all types. Production deployment is the highest-risk action in the AI lifecycle; multi-party authorization is required regardless of type to prevent unilateral deployment decisions. Quorum size may scale with risk type.
4.8.3	Dataset version approval digitally signed by data	SHALL	SHALL	SHALL	Uniform across all types. Prevents use of unapproved or untested

Commented [29]: "Actors = malicious actors"? Need definition

Commented [30]: Actors is qualified already. All Actors.



ReqID	Description	Type 1	Type 2	Type 3	Comments
	governance authority before use in training				data in training at every risk level; governance over data selection is foundational to model accountability.
4.8.4	Security updates and patches digitally signed by security team (unscheduled) or change management board (scheduled)	SHALL	SHALL	SHALL	Uniform across all types. Ensures all updates are authorized and traceable; prevents unauthorized code injection that could alter model behavior regardless of risk level.
4.8.5	Dataset updates/corrections digitally signed by data engineer (proposer) and data quality lead (approver)	SHALL	SHALL	SHALL	Uniform across all types. Two-party authorization prevents single-actor data manipulation at every risk level.
4.8.6	GDPR data subject requests (access Art. 15, erasure Art. 17) digitally signed by privacy officer (action) and legal (approval) with timestamp; resolution documented on-chain	SHALL	SHALL	SHALL	Uniform across all types. Creates an auditable compliance record for data protection authority investigations; GDPR applies regardless of AI risk type.
4.8.7	Performance anomaly acknowledgment and remediation plan digitally signed by on-call engineer (acknowledgment) and incident commander (remediation approval)	SHOULD	SHALL	SHALL	Type 1: Recommended for governance maturity but formal incident authorization is not a certification gate for non-critical systems. Type 2–3: Mandatory because anomalies in high-risk systems have



ReqID	Description	Type 1	Type 2	Type 3	Comments
					material consequences; signed acknowledgment ensures issues are formally recognized and not quietly dismissed, creating an unambiguous accountability record.
4.8.B – Technical Signature Requirements					
4.8.8	Signature algorithm: RSA (minimum 2048-bit; 3072- or 4096-bit recommended for Type 3), OR ECDSA (P-256, P-384, or P-521 curves), OR post-quantum (CRYSTALS-Dilithium, SPHINCS+, Falcon) for forward compatibility	SHALL	SHALL	SHALL	Uniform across all types — minimum algorithm requirements apply everywhere. Key length recommendations scale up at Type 3: post-quantum options are recommended for long-lived critical systems where cryptographic harvest-now-decrypt-later attacks are a concern. See Appendix C for an overview of quantum readiness relevant to AI assessors.
4.8.9	Public key published in verifiable key registry (on-chain registry, PKI certificate authority, or HSM-backed registry)	SHALL	SHALL	SHALL	Uniform across all types. Enables independent signature verification without relying on solution provider's claims at every risk level.
4.8.10	Each signature includes trusted timestamp (RFC 3161 Time-Stamp Protocol preferred, or	SHALL	SHALL	SHALL	Uniform across all types. Timestamps enable sequencing of actions and detection of



ReqID	Description	Type 1	Type 2	Type 3	Comments
	ISO 8601 timestamp attested by trusted time source)				retroactive authorization at every risk level.
4.8.C – Blockchain Recording					
4.8.11	Each signed action recorded as blockchain transaction: signer public key or identifier, action description, signature or hash of signature, ISO 8601 timestamp, reference to affected artifact (model-ID, dataset version, code commit hash)	SHALL	SHALL	SHALL	Uniform across all types. Creates immutable audit trail for post-incident forensics; the core accountability mechanism at every risk level.
4.8.12	Blockchain transactions linked to prior related transactions (e.g., deployment approval links to training approval, which links to dataset approval) forming a traceable chain of custody	SHOULD	SHALL	SHALL	Type 1: Recommended for governance maturity but hash-chaining overhead is not a certification requirement for non-critical systems. Type 2–3: Mandatory because hash-chaining enables full lineage reconstruction without manual document correlation — essential when investigating adverse outcomes or demonstrating regulatory compliance.
4.8.D – Authorization Matrix and Key Management					
4.8.13	Authorization matrix (RACI or equivalent) specifying: authorized	SHALL	SHALL	SHALL	Uniform across all types. Without an explicit authorization matrix,



ReqID	Description	Type 1	Type 2	Type 3	Comments
	actors by role, which actions require which actors' signatures, approval workflow, escalation paths				"all actors must sign" is unenforceable; the matrix operationalizes governance intentions at every risk level.
4.8.14	Key rotation schedule defined and documented (e.g., annual rotation; emergency rotation within 24 hours of suspected compromise)	SHOULD	SHALL	SHALL	Type 1: Recommended as key management best practice. Type 2–3: Mandatory because periodic rotation limits the blast radius of key compromise; for high-risk and critical systems, undetected key compromise could allow unauthorized model modifications with falsified audit trails.
4.8.15	Key revocation process documented; revocation list maintained on-chain or in verifiable registry with revocation date and reason (e.g., "employee termination", "key compromise")	SHOULD	SHALL	SHALL	Type 1: Recommended but low urgency for non-critical systems. Type 2–3: Mandatory because revoked or expired keys must be immediately unusable; for high-risk systems, a compromised signing key could be used to fraudulently authorize unauthorized changes, undermining the entire accountability framework.
4.8.16	Formal PKI or equivalent key management infrastructure	—	SHOULD	SHALL	Type 1: Not required; lightweight key management is



ReqID	Description	Type 1	Type 2	Type 3	Comments
	documented (key generation ceremony, storage in HSM or encrypted vault, distribution controls)				sufficient for non-critical systems. Type 2: Strongly recommended but formal PKI may be burdensome for smaller organizations; documented key management best practices are acceptable. Type 3: Mandatory because the severity of compromise in critical infrastructure demands hardware-backed key protection; informal key management creates unacceptable risk of unauthorized signing.

4.9 Model Output Visibility and Real-Time Monitoring

ReqID	Description	Type 1	Type 2	Type 3	Comments
4.9.A – Output Logging					
4.9.1	Detailed inference logs maintained per-prediction: ISO 8601 timestamp (millisecond precision), input features or SHA-256 hash thereof (if input is large or contains PII), model version that produced output,	SHOULD	SHALL	SHALL	Type 1: Recommended for monitoring capability but not a certification gate for low-risk systems. Type 2–3: Mandatory because per-prediction logs are the evidence base for investigating adverse outcomes; without them, it is impossible to



ReqID	Description	Type 1	Type 2	Type 3	Comments
	prediction/decision and confidence score				audit whether a specific high-stakes decision was appropriate or to identify systematic errors.
4.9.2	Actual outcome (ground truth) recorded when available, linked to corresponding prediction log entry, for retrospective accuracy and bias analysis	SHOULD	SHALL	SHALL	Type 1: Recommended for ongoing performance insight but ground truth collection is often not feasible for all low-risk applications. Type 2–3: Mandatory where ground truth is obtainable (financial outcomes, medical diagnoses, hiring outcomes); linkage to prediction logs enables ongoing validation against real-world performance, not just test set metrics.
4.9.3	Human actions on model outputs recorded: actor identity, any overrides or rejections with documented reason	SHOULD	SHALL	SHALL	Type 1: Useful for understanding human-AI interaction but not a certification gate for non-critical systems. Type 2–3: Mandatory because separating "model recommendation" from "human decision" is essential for accountability; override patterns reveal whether human oversight is

Commented [31]: "actor" defined?



ReqID	Description	Type 1	Type 2	Type 3	Comments
					functioning or whether humans are rubber-stamping all model outputs.
4.9.4	Log retention enforced: Type 1 minimum 6 months; Type 2 minimum 1 year (extend to 5–7 years for financial services/healthcare per sector regulation); Type 3 minimum 2 years or as required by applicable law	SHALL	SHALL	SHALL	Retention minimum differs by type reflecting regulatory exposure: Type 1 logs are for operational monitoring only; Type 2 logs must support regulatory investigations and potential litigation; Type 3 logs must support long-term forensic analysis and may be subject to sector-specific retention laws (e.g., aviation safety logs: 10 years).
4.9.B – Blockchain Recording of Outputs					
4.9.5	Type 1 sampling approach: 5–10% of outputs sampled (random or stratified) with hashes of sampled outputs recorded on-chain with timestamp and model version reference	SHOULD	—	—	Type 1 only: Representative sampling is sufficient for spot-check auditability of non-critical systems without incurring high on-chain storage costs. Type 2–3: Not applicable; aggregate statistics (4.9.6) replace sampling for higher-risk systems to provide more comprehensive coverage.



ReqID	Description	Type 1	Type 2	Type 3	Comments
4.9.6	Type 2–3 aggregate statistics approach: Output statistics recorded on-chain daily or weekly: total predictions, output distribution (e.g., approval/denial rates), performance metrics if ground truth available, anomalies or threshold breaches flagged	—	SHALL	SHALL	Type 1: Not applicable; sampling (4.9.5) is appropriate for low-risk systems. Type 2–3: Mandatory because aggregate on-chain recording provides accountability at scale without recording millions of individual transactions; combined with off-chain detail logs (4.9.7), this achieves both scalability and auditability.
4.9.7	Off-chain detailed inference logs linked to on-chain aggregate via hash (auditors verify log authenticity by computing hash of log file and comparing to on-chain record)	SHOULD	SHALL	SHALL	Type 1: Recommended for log integrity assurance but not a certification gate. Type 2–3: Mandatory because the hash link is what makes the off-chain logs auditable; without it, logs could be altered retroactively and the on-chain aggregate would not detect the manipulation.
4.9.C – Performance Monitoring and Alerting					
4.9.8	Production performance and drift monitoring implemented. The solution shall monitor applicable primary performance metrics, input distribution drift,	SHOULD	SHALL	SHALL	Consolidates the common monitoring and alerting elements formerly repeated in 4.9.10 and 4.9.11. Fairness monitoring



ReqID	Description	Type 1	Type 2	Type 3	Comments
	and output distribution drift; define thresholds; and trigger alerts when thresholds are breached.				remains separately addressed in 4.9.9.
4.9.9	Automated monitoring of fairness metrics in production (demographic parity, equalized odds, calibration by group); alert triggered if fairness metric exceeds threshold (e.g., demographic parity difference > 0.05)	—	SHALL	SHALL	Type 1: Not required; fairness monitoring overhead is disproportionate for non-consequential decisions. Type 2–3: Mandatory because bias can emerge in production even when pre-deployment training metrics were acceptable (data drift causes differential performance shifts); continuous fairness monitoring is necessary for systems whose decisions affect protected groups.
4.9.10	Input-distribution drift detection implemented as an evidence element of the production performance and drift monitoring control in 4.9.8 (e.g., Kolmogorov-Smirnov, chi-squared, or Population Stability Index).	SHOULD	SHALL	SHALL	Cross-reference added; applicability remains unchanged.
4.9.11	Output-distribution drift detection implemented as an evidence element of the production	SHOULD	SHALL	SHALL	Cross-reference added; applicability remains unchanged.



ReqID	Description	Type 1	Type 2	Type 3	Comments
	performance and drift monitoring control in 4.9.8, with an alert on unexpected change in prediction distribution.				
4.9.12	Alert-to-resolution workflow implemented. Each material alert shall record the timestamp, metric, threshold, current value, model version, assigned escalation recipient, response actions, root cause or false-alarm determination, corrective measure, and closure status in a tamper-evident incident-management record or blockchain transaction.	SHOULD	SHALL	SHALL	Consolidates the alert-recording and resolution documentation formerly repeated in 4.9.14. Escalation timeframes remain specified in 4.9.13.
4.9.13	Escalation protocol enforced: Type 1 alert (minor) → ML engineer within 24 hours; Type 2 alert (moderate) → ML lead within 4 hours, remediation plan within 5 business days; Type 3 alert (critical) → immediate incident commander escalation, potential rollback, root cause analysis, remediation plan within 24 hours	SHOULD	SHALL	SHALL	Type 1: Recommended response structure for operational discipline. Type 2–3: Mandatory because without tiered escalation, all alerts are treated equivalently and critical issues may be deprioritized; the escalation ladder ensures response urgency matches consequence severity.



ReqID	Description	Type 1	Type 2	Type 3	Comments
4.9.14	Alert resolution shall be documented as an evidence element of the alert-to-resolution workflow required by 4.9.12.	SHOULD	SHALL	SHALL	Cross-reference added to remove duplicate closure-documentation language.
4.9.D – Output Privacy Protection					
4.9.15	If outputs contain or reveal PII: outputs stored as cryptographic hashes on-chain only; full outputs held encrypted off-chain, accessible to auditors via secure facility or encrypted transfer	SHALL	SHALL	SHALL	Uniform across all types. Protects data subject privacy while maintaining auditability; required regardless of type because GDPR applies to any system processing personal data.
4.9.16	Aggregated performance reports anonymized before public disclosure; individual decision records not publicly disclosed	SHALL	SHALL	SHALL	Uniform across all types. Public transparency on aggregate performance without exposing individual decisions; maps to GDPR data minimization principle regardless of AI risk level.
4.9.17	Differential privacy applied to public statistical reports where individual re-identification risk is non-trivial (ϵ -DP or δ -DP; privacy budget documented)	—	SHOULD	SHOULD	Type 1: Not required; the simpler aggregation and anonymization measures in 4.9.16 are sufficient for low-risk systems. Type 2–3: Strongly recommended where DPIA identifies re-identification risk from published aggregate statistics; differential



ReqID	Description	Type 1	Type 2	Type 3	Comments
					privacy provides formal mathematical guarantees that individual records cannot be inferred from published summaries.

4.10 Additional Requirements for High-Risk and Critical-Infrastructure Solutions

Commented [32]: Replace with : "High-risk AI systems (Annex III)" if referring to EU AI Act.

ReqID	Description	Type 1	Type 2	Type 3	Comments
4.10.A – Hardware Compatibility					
4.10.1	Hardware specifications and constraints documented: minimum CPU, RAM, GPU/TPU requirements, storage requirements, peak memory usage during inference	—	SHOULD	SHALL	Type 1: Not required; hardware documentation overhead is disproportionate for non-critical systems. Type 2: Recommended so that deployers understand resource requirements, but not a certification gate. Type 3: Mandatory because critical systems must demonstrate hardware-level resource guarantees; under-provisioned hardware in a safety-critical deployment (e.g., autonomous vehicle inference latency exceeding safe



ReqID	Description	Type 1	Type 2	Type 3	Comments
					response time) can cause direct harm.
4.10.2	Computational performance benchmarked and documented: inference latency (ms/prediction), throughput (predictions/second), behavior under peak load	—	SHOULD	SHALL	Type 1: Not required for non-critical systems. Type 2: Recommended to inform deployment planning. Type 3: Mandatory because safety-critical systems must meet latency requirements under all operating conditions, including peak load; performance degradation under load in a critical system (e.g., medical diagnosis, air traffic control) may be life-threatening.
4.10.3	Compatibility with target deployment environments documented (cloud: AWS/Azure/GCP; on-premises; edge: mobile/IoT) including degraded performance modes	—	SHOULD	SHALL	Type 1: Not required for non-critical systems. Type 2: Useful deployment planning documentation but not a certification gate. Type 3: Mandatory because deploying a critical system into an unsupported environment where performance is



ReqID	Description	Type 1	Type 2	Type 3	Comments
					unknown or degraded creates unacceptable safety risk; compatibility must be verified and documented before deployment.
4.10.B – Human Oversight					
4.10.4	Meaningful human oversight established and documented. For applicable decisions, the solution shall define and enforce: (a) proportion or categories of decisions subject to human review; (b) thresholds requiring human approval; (c) a human override capability with auditable logging; and (d) training and qualification requirements for reviewers.	—	SHOULD	SHALL	Consolidates the operational elements of human oversight formerly distributed across 4.10.4–4.10.7. Type-specific applicability is retained through the supporting rows.
4.10.5	Human-approval thresholds shall be documented as an evidence element of the meaningful human oversight control in 4.10.4.	—	SHOULD	SHALL	Cross-reference added to remove overlap while retaining the applicable evidence element and risk-type designation.
4.10.6	Human override capability and auditable override logs shall be implemented as an evidence element of the meaningful human	—	SHALL	SHALL	Cross-reference added to remove overlap while retaining the applicable evidence



ReqID	Description	Type 1	Type 2	Type 3	Comments
	oversight control in 4.10.4.				element and risk-type designation.
4.10.7	Reviewer training and qualification requirements shall be defined as an evidence element of the meaningful human oversight control in 4.10.4.	—	SHOULD	SHALL	Cross-reference added to remove overlap while retaining the applicable evidence element and risk-type designation.
4.10.C – Fairness and Bias Mitigation					
4.10.8	Fairness metrics selected and justified for the use case and applicable legal framework (demographic parity, equalized odds, equal opportunity, calibration by group)	—	SHOULD	SHOULD	Type 1: Not required for non-critical systems. Type 2–3: Strongly recommended because different fairness metrics can be mutually incompatible (satisfying one may violate another); selection must be justified based on the specific use case, affected population, and applicable anti-discrimination law — there is no universally correct fairness metric.
4.10.9	Pre-deployment bias testing conducted across all relevant demographic groups; results	—	SHALL	SHALL	Type 1: Not required for non-critical systems. Type 2–3: Mandatory because high-risk systems



ReqID	Description	Type 1	Type 2	Type 3	Comments
	documented and available to auditors				cannot be deployed without documented evidence of differential performance across groups; deploying without bias testing means accepting unknown discriminatory outcomes — unacceptable for decisions affecting individual rights or safety.
4.10.10	Debiasing techniques applied where bias is detected (data rebalancing, fairness-constrained optimization, adversarial debiasing); technique choice justified	—	SHOULD	SHALL	Type 1: Not required. Type 2: Strongly recommended when bias is detected; technique choice depends on the fairness metric, model architecture, and acceptable accuracy trade-offs — justification is required to demonstrate that the chosen approach is appropriate and effective for the specific context. Type 3: Mandatory because any bias can cause problems down the



ReqID	Description	Type 1	Type 2	Type 3	Comments
					road and lead to sub-optimal outcomes.
4.10.11	Ongoing fairness monitoring in production linked to formal fairness governance: designated Fairness Officer or equivalent, quarterly fairness audit reports, defined escalation if in-production bias detected	—	SHOULD	SHALL	Type 1: Not required. Type 2: Strongly recommended for maturing governance programs. Type 3: Mandatory because fairness can degrade post-deployment as the served population diverges from the training distribution; without formal governance, in-production bias may persist undetected and unaddressed for extended periods in critical applications.
4.10.D – Explainability and Interpretability					
4.10.12	Human-readable explanations provided to affected individuals for high-consequence decisions (e.g., "Loan denied because debt-to-income ratio 48% exceeds maximum threshold 43%")	—	SHOULD	SHALL	Type 1: Not required for non-critical systems. Type 2: Strongly recommended; required by GDPR Art. 22 for fully automated decisions with legal or similarly significant effect — assess whether this applies. Type 3: Mandatory because affected individuals have a legal right to



ReqID	Description	Type 1	Type 2	Type 3	Comments
					understand decisions affecting their safety, health, or fundamental rights; required by EU AI Act Art. 13 transparency obligations for high-risk AI.
4.10.13	Feature importance explanations available to auditors and authorized stakeholders (SHAP, LIME, attention weights, saliency maps, or equivalent); method documented and justified	—	SHOULD	SHALL	Type 1: Not required for non-critical systems. Type 2: Strongly recommended to enable auditors to verify the model is not using proxy variables for protected characteristics. Type 3: Mandatory because assessors must be able to verify that safety-critical and high-stakes models are relying on appropriate features, not spurious correlations; feature importance is the primary tool for detecting hidden bias and model shortcuts.
4.10.14	Counterfactual explanations available for adverse decisions (e.g., "If income were \$42,000	—	SHOULD	SHOULD	Type 1: Not required. Type 2–3: Strongly recommended as the most actionable form of explanation for



ReqID	Description	Type 1	Type 2	Type 3	Comments
	[+\$7k], decision would change to Approved")				affected individuals, enabling informed appeals and understanding of what would change the outcome. Not mandatory because counterfactual generation is technically challenging for complex models and may not always be feasible.
4.10.15	Model card published documenting: capabilities, limitations, fairness characteristics, intended uses, prohibited uses, training data summary, performance metrics across demographic groups	—	SHOULD	SHALL	Type 1: Not required. Type 2: Strongly recommended as a standardized transparency artifact enabling consistent stakeholder and procurement assessment. Type 3: Mandatory because critical system operators, downstream integrators, and regulators require standardized, comprehensive model documentation; ad hoc documentation is insufficient for systems where failure has severe consequences.



ReqID	Description	Type 1	Type 2	Type 3	Comments
4.10.E – Incident Response and Forensics					
4.10.16	AI incident response, investigation, and remediation process documented and operational. The process shall define notification and escalation, response timelines, containment or rollback authority, root-cause analysis, corrective-action tracking, and time-bound remediation.	—	SHOULD	SHALL	Consolidates the incident lifecycle formerly distributed across 4.10.16–4.10.18. Type-specific applicability is retained through the supporting rows.
4.10.17	Post-incident review and root-cause analysis shall be performed as an evidence element of the incident response, investigation, and remediation process required by 4.10.16.	—	SHOULD	SHALL	Cross-reference added to remove overlap while retaining the applicable evidence element and risk-type designation.
4.10.18	Time-bound remediation and corrective-action tracking shall be performed as an evidence element of the incident response, investigation, and remediation process required by 4.10.16.	—	SHOULD	SHALL	Cross-reference added to remove overlap while retaining the applicable evidence element and risk-type designation.

5 Assessment Methodology and Compliance Determination

5.1 Overview

This section describes how solutions are assessed for compliance with the AI Supplement requirements and how certification determinations are made. The assessment process follows the



GBA BMM assessment framework (see BMM Assessment Requirements document) with AI-specific adaptations.

5.2 Assessment Team Composition

A BMM AI assessment team shall consist of at least 4 qualified members:

1. **AI Security/ML Subject Matter Expert (AI SEC SME):** Qualified per Section 3 training pathway (100-hour curriculum or waiver)
2. **Domain Expert:** Expert in the solution's application domain (healthcare, finance, supply chain, manufacturing, etc.)
3. **AI Regulatory/Compliance Expert:** (GDPR, EU AI Act, ISO 42001, NIST RMF, etc.) depending on the jurisdiction in which the regulations are in effect.
4. **Lead Assessor:** A current GBA-qualified lead assessor responsible for assessment governance, conflict resolution, final determination, and report sign-off. The Lead Assessor may also satisfy one of the other competency roles where qualified.

The combined team **must have expertise** in:

- **AI systems:** Architecture, training methodologies, common failure modes, adversarial attacks
- **Blockchain/distributed ledgers:** Consensus mechanisms, smart contracts, cryptographic verification, on-chain vs. off-chain storage
- **Data privacy and protection:** GDPR compliance, data minimization, anonymization techniques, right-to-erasure accommodation
- **Applicable regulations:** For the solution's domain (e.g., FDA regulations for medical AI, FINRA for financial AI)
- **Assessment methodology:** Evidence collection, documentation review, consensus-building, findings documentation

Team Independence:

- **At least 50% of team members must be external** to the organization seeking assessment (maintains objectivity)
- **No team member may have supervisory authority** over another team member (prevents conflicts of interest)
- **No consultants who advised on solution design** may serve on assessment team for that solution (separation of consulting and auditing)



5.3 Evidence Collection and Review

Assessors **shall review evidence** that the solution satisfies AI Supplement requirements. Evidence types include:

Documentation:

- Architecture documents (system design, data flow diagrams, deployment architecture)
- Design specifications (model architecture, training methodology, hyperparameters)
- Process documentation (governance policies, change management procedures, incident response plans)
- Risk management protocols (Section 4.4.2 requirements)

Artifacts:

- Source code (Section 4.2): Actual code or hash **verification** if conditional disclosure
- Model files (Section 4.3): Model binaries or hash verification
- Training scripts and configuration files
- Dataset metadata (Section 4.5): Provenance, composition, version control

Audit Trails:

- Blockchain transaction history (all lifecycle events, signatures, approvals)
- Digital signatures verification (Section 4.8): Verify signatures cryptographically against public keys
- Approval logs and change management records**

Test Results:

- Model validation reports (Section 4.6: Training loop documentation, performance metrics)
- Adversarial robustness testing results (if applicable)
- Fairness assessments (Section 4.10.3: Bias testing across demographic groups)
- Performance monitoring reports (Section 4.9.3: Production metrics, drift detection)

Interviews:

- Development team (data scientists, ML engineers): "Walk us through your training process"
- Compliance officers: "How do you ensure GDPR compliance?"
- Deployment engineers: "How do you monitor model performance in production?"
- Security team: "What adversarial attack mitigations are in place?"
- General: "What questions should we have asked you but did not?" – designed to elicit information that the other questions might have missed.

Demonstrations:

- Live system walkthrough (show model inference, output logging, monitoring dashboards)
- Hash verification (assessor computes hash of provided artifact, verifies against on-chain record)

Commented [33]: Add prose to explicitly mention formal code verification techniques. Section 4.2 might be the best location for this prose.



- Signature verification (assessor verifies digital signatures using public keys)
- Output logging verification (assessor triggers test prediction, verifies it appears in logs)

5.4 Risk-Type Determination

Before detailed requirement-by-requirement assessment, the team **shall determine the solution's risk type** (Type 1, 2, or 3) based on Section 1.2 criteria. This determination guides which requirements are mandatory vs. recommended.

Decision Factors:

- **Impact of AI decisions:** Are decisions binding/irreversible (e.g., deny loan, deny insurance) or advisory/reversible (e.g., content recommendation that user can ignore)?
- **Scope of deployment:** Single organization internal use, or broad market deployment affecting thousands/millions of users?
- **Critical infrastructure elements:** Does failure pose risk to life, safety, national security, or critical infrastructure?
- **Affected populations:** Small user base (e.g., internal employees) or large, potentially vulnerable populations (e.g., consumers, patients, students)?
- **Regulatory classification:** Is the solution classified as "high-risk" under EU AI Act or similar regulation?

Risk Type Decision Tree:

Q1: Does the AI system make decisions that could materially affect individual rights, safety, or access to essential services?

└─ NO → **TYPE 1 (Low-Risk)**

└─ YES → Q2

Q2: Are decisions binding/irreversible or reversible/advisory?

└─ Reversible/Advisory → **TYPE 1 (Low-Risk)** [unless Q3 elevates]

└─ Binding/Irreversible → Q3

Q3: Is this critical infrastructure, safety-critical, or high-impact domain (healthcare diagnostics, autonomous vehicles, judicial, military)?



NO → TYPE 2 (High-Risk)

YES → TYPE 3 (Critical Infrastructure)

Commented [34]: widow orphan control. keep no and yes blocks on same page

Examples:

- Type 1: Netflix content recommendation, inventory optimization, sentiment analysis
- Type 2: Loan approval, hiring decision, insurance underwriting, content moderation
- Type 3: Medical diagnosis support, autonomous driving, air traffic control, nuclear plant monitoring

Risk Type Implications for Assessment:

The applicability designations in the Section 4 requirement tables control the assessment. Assessors shall determine compliance requirement by requirement, using the **SHALL**, **SHOULD**, **MAY**, and **Not Applicable** designations shown for the solution's assigned risk type. Section-level summaries are descriptive only and do not override an individual requirement's applicability designation.

- **Type 1 (Low-Risk AI):** Type 1 assessments emphasize baseline transparency, documentation, provenance, accountability, privacy, and lifecycle controls designated **SHALL** in the Section 4 tables. Additional controls are designated **SHOULD** where they represent recommended practices proportionate to a low-risk deployment. Some controls are **MAY** or **Not Applicable** where they are not proportionate to the solution's risk and context. A Type 1 system is not required to satisfy every requirement in Sections 4.2 through 4.9; it is required to satisfy each Type 1 **SHALL** requirement. For example, source-code provenance, architecture documentation, deployment documentation, known limitations, model documentation, dataset provenance, privacy protections, training-run audit records, actor signatures, and specified output-privacy controls remain mandatory where marked **SHALL**.
- **Type 2 (High-Risk AI):** Type 2 assessments require compliance with all Section 4 requirements designated **SHALL** for Type 2. In practice, this includes most documentation, provenance, data-governance, training, model-integrity, accountability, and production-monitoring controls in Sections 4.2 through 4.9. However, Type 2 does not make every listed requirement mandatory: some requirements remain **SHOULD**, **MAY**, or **Not Applicable** because they depend on the system's architecture, deployment model, data sensitivity, or operational context. For example, public source-code release is a **SHOULD**, controlled access to source code is a **SHALL**, hash-only source-code disclosure is not permitted, and certain advanced privacy or resilience measures remain recommended rather than mandatory.
- **Type 3 (Critical-Infrastructure AI):** Type 3 assessments require compliance with all Section 4 requirements designated **SHALL** for Type 3, including the applicable enhanced controls in



Section 4.10. This generally results in the most extensive assessment scope: provenance and lifecycle records; source-code or alternative controlled assurance; SBOM and supply-chain controls; model, dataset, and training documentation; reproducibility evidence; secure model-weight management; authorization and key management; detailed output logging and monitoring; meaningful human oversight; fairness, explainability, and model-card controls where applicable; and incident response, investigation, and time-bound remediation. Certain controls may nevertheless remain **SHOULD**, **MAY**, or **Not Applicable** for Type 3 where public disclosure, a specific technical technique, or a deployment-specific condition is not appropriate. For example, public source-code or model-weight release may be optional or inappropriate for classified or safety-sensitive systems, while controlled assessor access and meaningful independent assurance remain required.

Assessment rule: A solution cannot receive the requested BMM AI certification level if a requirement designated **SHALL** for its assigned risk type is non-compliant. A **SHOULD** or **MAY** requirement is not a basis for a non-compliant certification finding, although assessors may identify an unmet **SHOULD** requirement as an observation, improvement opportunity, or recommended remediation.

5.5 Compliance Determination

For each requirement in Section 4 applicable to the solution's risk type, the team shall determine:

- **Compliant:** Evidence clearly demonstrates solution satisfies requirement fully. No gaps or weaknesses identified.
- **Partially Compliant:** Solution addresses most but not all aspects of requirement. Minor gaps exist but do not fundamentally undermine control.
 - Example: "Source code is available to auditors under NDA (Option B), but hash is not yet recorded on-chain. Gap: Missing on-chain hash commitment. Remediation: Record hash within 30 days."
- **Non-Compliant:** Solution does not address requirement, or gaps are material and undermine the control's effectiveness.
 - Example: "No performance monitoring in production. Model deployed 6 months ago with no drift detection. Non-compliant with Section 4.9.8."
- **Not Applicable:** Requirement does not apply to this solution's risk type or domain context.
 - Example: "Section 4.10.4 (human oversight) is **SHOULD** for Type 2; solution chose not to implement. Marked N/A for certification purposes (not required for Type 2)."

Consensus Process:



- Assessment team **must reach consensus** on compliance determination for each requirement
- If disagreement exists, team discusses evidence and rationale until consensus achieved
- If consensus cannot be reached, **Lead Assessor makes final determination** (with dissenting opinions documented)
- All determinations **must be evidence-based** (no subjective opinions without supporting evidence)

5.5.1 Partial vs. Non-Compliant Determination Rule (mandatory)

A finding of Partially Compliant requires that all of the following hold: (a) the control's core assurance objective is substantively met; (b) the identified gap is narrow and discrete, and does not by itself create a material risk of harm, fraud, unauthorized access, or undetected failure; (c) the gap is remediable without redesigning the control; and (d) the solution provider has committed to a documented remediation plan with a bounded deadline. If any condition is not met, the finding shall be Non-Compliant.

5.5.2 Non-Waivable Controls (mandatory)

The following categories of SHALL requirements shall never receive a Partially Compliant finding; they are Compliant or Non-Compliant only, because a partial gap creates unacceptable risk: cryptographic signature and key-management controls (Section 4.8); artifact provenance and lifecycle-integrity controls (Sections 4.2.4, 4.3.7, 4.3.11, 4.5.7, 4.7.1, 4.6.20); GDPR privacy and erasure controls (Section 4.5.9–4.5.13); Type 3 meaningful human-oversight controls (Section 4.10.4); and Type 3 incident-response controls (Section 4.10.16). All other SHALL requirements remain eligible for Partially Compliant under 5.5.1.

5.5.3 Remediation Deadlines and Certification Status (mandatory)

Where a Partially Compliant finding is made on an eligible requirement, certification is issued as Provisional Certification, valid for the remediation window: 30 days for a major gap on a Type 3 system, 60 days for a major gap on Type 1/2 or a minor gap on Type 3, and 90 days for a minor gap on Type 1/2. Provisional Certification converts automatically to full certification upon verified remediation. If the deadline passes without verified remediation, or if any Non-Compliant finding exists on a non-waivable control under 5.5.2, certification at the requested level shall be withheld until remediation is verified.



5.6 AI Supplement Rating and Certification Level

AI Supplement Rating Scale:

The AI Supplement uses a 5-level maturity scale aligned with the baseline BMM:

- **Level 1 (Initial):** Basic transparency and documentation exist. Minimal governance structure. Some AI artifacts available but incomplete. Audit trail sporadic.
 - Evidence: Model architecture documented; source code available; dataset metadata published; no formal governance
- **Level 2 (Documented):** Full documentation of all AI artifacts. Formal governance policies defined. Blockchain audit trail for key lifecycle events (training, deployment). Risk management protocols documented.
 - Evidence: All requirements in Sections 4.2–4.5 satisfied; governance policies approved; training audit trail on-chain
- **Level 3 (Validated):** Validated training and testing procedures. Real-time monitoring of model performance. Fairness testing conducted. Digital signatures on all material actions. GDPR compliance mechanisms implemented.
 - Evidence: All requirements in Sections 4.2–4.8 satisfied; production monitoring operational; fairness metrics tracked
- **Level 4 (Production):** Production-grade governance with continuous monitoring. Incident response plan operational. Performance alerting and escalation functional. Human oversight mechanisms in place (if applicable). Post-deployment evaluation regular.
 - Evidence: All requirements in Sections 4.2–4.9 satisfied; incident response tested; continuous improvement process documented
- **Level 5 (Optimizing):** Continuous improvement culture. Advanced techniques (federated learning, differential privacy, adversarial training) implemented. Proactive threat detection. Multi-stakeholder governance. Regular third-party audits.
 - Evidence: All requirements satisfied; advanced privacy-preserving techniques deployed; external red-team testing conducted**

Certification Criteria:

A solution may be rated at BMM AI Level X only if it satisfies:

- All baseline BMM element requirements at Level X (all 11 BMM elements: Distribution, Governance, Identity Management, Interoperability, Performance, Privacy, Reliability, Resilience, Security, Sustainability, Synchronization), **AND**
- All AI Supplement requirements applicable to the solution's risk type (as determined in Section 5.4), **AND**
- No non-compliant findings on mandatory ("shall") requirements for that risk type



Partial Compliance and Remediation:

- If any mandatory requirement is **non-compliant**, certification is not awarded at that level
- Solution may be recommended for Level X-1 (if lower-level requirements satisfied)
- Or certification deferred pending remediation

If **Partially Compliant** findings exist, certification **may be conditional:**

- "Certified at Level 3 conditional on remediation of [specific gap] within 90 days"
- Follow-up review required to confirm remediation
- **Remediation plan required** for all Partially Compliant and Non-Compliant findings:
 - What will be fixed?
 - Who is responsible?
 - What is the timeline?
 - How will compliance be verified?

Commented [35]: Get Gerard's reaction to the conditional certification. A simpler option might be to just defer assessment until remediation is done.

Example Certification Determination:

Solution: AI-Powered Loan Underwriting System

Risk Type: Type 2 (High-Risk) — decisions affect individual rights (access to credit)

Baseline BMM: Level 3 (Validated) — all 11 elements satisfied at Level 3

AI Supplement Assessment:

- Section 4.2 (Source Code): Compliant (Option B: conditional disclosure to auditors)
- Section 4.3 (Model Availability): Compliant (Option A: model binary available via secure download)
- Section 4.4 (Open-Source Licensing): Compliant (proprietary; risk management protocols published)
- Section 4.5 (Dataset): Compliant (metadata published; hybrid on-chain/off-chain GDPR accommodation)
- Section 4.6 (Training Loop): Compliant (full documentation; training audit trail on-chain)
- Section 4.7 (Model Weights): Compliant (hash on-chain; weights available to auditors)
- Section 4.8 (Signatures): Compliant (all material actions signed; PKI documented)
- Section 4.9 (Monitoring): Partially Compliant — aggregate statistics recorded on-chain, but alerting thresholds not yet configured

Gap: Monitoring dashboard operational but no automated alerts



Remediation: Configure alerts for accuracy <90%, demographic parity >0.05; deploy within 60 days

Certification Determination:

Certified at BMM Level 3 for AI (Type 2), conditional on completion of monitoring alert configuration within 60 days. Follow-up review required to verify remediation.

5.7 Assessment Report and Recommendations

The assessment shall produce a written report including:

Report Sections:

1. **Executive Summary** (1–2 pages)
 - Solution name, version, vendor/provider, domain
 - Risk type determination and rationale
 - Overall certification determination (Level X, conditional, deferred, denied)
 - High-level strengths and key gaps
2. **Assessment Scope and Methodology** (1 page)
 - Assessment date(s) and location (on-site, remote)
 - Assessment team members and qualifications
 - Evidence reviewed (documentation, artifacts, interviews, demonstrations)
3. **Risk Type Determination** (1 page)
 - Decision factors evaluated (impact, scope, criticality, affected populations)
 - Risk type assigned (Type 1, 2, or 3) with justification
 - Applicable requirements (which sections of Section 4 are mandatory vs. recommended)
4. **Requirement-by-Requirement Findings** (10–20 pages, depending on complexity)
 - For each requirement: Compliance determination (Compliant, Partially Compliant, Non-Compliant, N/A)
 - Evidence cited (document references, interview notes, demonstration results, artifact verification)
 - Gaps identified (for Partially Compliant and Non-Compliant findings)
 - Strengths highlighted (best practices, innovative approaches)
5. **Overall Strengths and Best Practices** (1–2 pages)
 - What is the solution doing exceptionally well?
 - Innovative approaches worth highlighting
 - Exemplary governance or technical practices
6. **Gaps and Improvement Opportunities** (2–3 pages)



- Summary of all Partially Compliant and Non-Compliant findings
 - Prioritized by severity (critical, major, minor)
 - Impact of each gap on overall trustworthiness and governance
7. **Certification Determination** (1 page)
- Recommended AI Supplement rating (Level 1, 2, 3, 4, or 5)
 - Conditions or caveats (if applicable): "Certified conditional on remediation of [gap] within [timeline]"
 - Limitations or exclusions: "Fitness for healthcare domain not assessed; certification applies to financial services use case only"
8. **Appendices**
- Assessment checklist (requirement-by-requirement evidence matrix)
 - Evidence log (documents reviewed, artifacts verified, interviews conducted)
 - Blockchain transaction verification log (hashes verified, signatures validated)
 - Team member qualifications and conflict of interest disclosures

Report Distribution:

- **Solution provider** receives full report
- **GBA Director of Standards & Certification** receives report for review and approval
- **Public summary** published on GBA website (if solution provider consents): Executive summary + certification determination only (detailed findings confidential)

Appendix A: Glossary

The following terms are used in this supplement with specific meanings. Terms defined in the BMM Overview (v2.0) are incorporated by reference.

Term	Definition
Artificial Intelligence (AI) Model	A mathematical representation or computational system trained on data to make predictions, classifications, or decisions. Includes neural networks, decision trees, ensemble methods, large language models (LLMs), and other machine learning architectures.
AI Risk Type	Classification of AI system's risk profile based on potential impact of decisions and deployment scope. Type 1: Low-Risk (non-critical); Type



Term	Definition
	2: High-Risk (affects individual rights, financial decisions); Type 3: Critical Infrastructure (safety-critical, life-critical, national security). Determines which requirements are mandatory ("shall") vs. recommended ("should").
AI Security Assessment Subject Matter Expert (AI SEC SME)	Assessment team member qualified to evaluate AI systems per Section 3 training pathway. Must complete 100-hour curriculum (or equivalent via waiver) covering AI fundamentals, security, blockchain auditability, regulation, and assessment methodology.
Audit Trail	Immutable, time-ordered record of actions, decisions, and changes to AI system artifacts (data, models, code, deployments). Recorded on blockchain for tamper-proof verification and forensic analysis.
Bias and Fairness	Bias: Systematic disparity in model predictions or error rates across demographic groups (gender, race, age, disability status, etc.). Fairness: Property that model predictions satisfy equity criteria such as demographic parity (equal approval rates across groups), equalized odds (equal false positive/negative rates), or calibration (predicted probabilities match observed outcomes across groups). Mitigated through fairness-aware training, data rebalancing, and post-processing adjustments.
Data Poisoning	Adversarial attack where attacker manipulates training data to cause incorrect AI behavior. Examples: label flipping (changing labels from correct to incorrect), backdoor insertion (injecting trigger patterns causing specific misclassifications), gradient-based attacks. Mitigated through data validation, anomaly detection, and robust training techniques.
Differential Privacy	Mathematical framework providing formal privacy guarantees (ϵ -differential privacy). Ensures aggregated statistics about a dataset do not reveal information about individual records. Enables public reporting of model performance metrics while protecting individual privacy. Implemented by adding calibrated statistical noise to query results.



Term	Definition
Explainability / Interpretability	Ability to understand and explain AI model decisions in human-comprehensible terms. Interpretability: Model's internal logic is transparent (e.g., decision trees, linear models). Explainability: Post-hoc techniques explain black-box model decisions (e.g., SHAP, LIME, attention visualization). Critical for high-risk decisions requiring accountability and auditability.
Hash Commitment	Cryptographic hash (SHA-256 or stronger) of an artifact (dataset, model, source code) recorded on blockchain. Provides tamper-proof proof of existence and integrity at specific timestamp. Enables verification without exposing artifact content (privacy-preserving transparency).
Large Language Models (LLMs)	Large Language Models (LLMs): Neural network models trained on massive text corpora (billions of tokens) to generate human-like text. Examples: GPT-4, Claude, PaLM, LLaMA. Special considerations: prompt injection attacks, hallucinations, alignment challenges.
Model Drift	Degradation in model performance over time due to: • Data drift: distribution of input features changes (e.g., economic conditions shift, user behavior evolves) • Concept drift: relationship between inputs and outputs changes (e.g., definition of "creditworthy" evolves) • System drift: changes in data collection or preprocessing. Detected through continuous monitoring (Section 4.9.3). Mitigated through retraining or model updating.
Off-Chain Storage	Data stored outside the blockchain (encrypted database, IPFS, secure cloud storage, data warehouse). Maintains mutable, access-controlled storage for large datasets or personal data. Complements on-chain hash commitments (blockchain stores hashes for integrity verification; off-chain storage holds actual data).
On-Chain Storage	Data recorded directly on the blockchain as transactions or smart contract state. Immutable (cannot be deleted or modified retroactively). Used for audit trails, hash commitments, digital



Term	Definition
	signatures, and compliance documentation. Limited by blockchain storage capacity (expensive for large data).
Stakeholder (Relevant)	Person or entity with specific role in AI solution lifecycle (developer, data scientist, governance board member, deployer) or materially affected by AI decisions (end user, data subject, regulator).
Stakeholder (General)	Any person or entity that may be impacted by the AI solution's decisions, performance, or security. Includes users, customers, employees, regulators, auditors, civil society organizations, affected communities.
Tiered Disclosure	Risk-based approach to transparency with different access levels for different stakeholder types. ● Public disclosure (metadata, performance metrics, risk management protocols), ● Conditional disclosure (source code/models to auditors under NDA), ● Minimal disclosure (hash commitment only for national security/IP justification). Balances transparency with legitimate confidentiality needs.
Transparency (AI Transparency)	Property that stakeholders can access AI system documentation, source code, models, training data metadata, and decision rationale sufficient for understanding and informed decision-making. Enables auditing, accountability, and trust. Implemented through artifact publication (Section 4.2–4.5) and blockchain-backed hash commitments.



Appendix B: Mapping to Global AI Frameworks

This appendix maps AI Supplement requirements to major international AI governance frameworks, demonstrating alignment and facilitating organizations' compliance with multiple standards, wherever compliance with such standards is mandatory.

AI Supplement Requirement	NIST AI RMF 1.0	EU AI Act (Reg 2024/1689)	ISO/IEC 42001:2023	OECD AI Principles
Section 4.2: Source Code Transparency	MAP-2.1: Document AI system design, components, and development process	Annex IV §2(a): Technical documentation including system design and development	Clause 7.2.3: Documentation of AI system development	Transparency: Enable understanding of AI system outcomes
Section 4.3: Model Availability & Versioning	MEASURE-2.1: Validate and test AI system; document performance	Annex IV §2(f): Training, validation, testing methodologies and results	Clause 7.2.5: AI system validation and performance evaluation	Accountability: Enable tracing of AI system outcomes
Section 4.5: Dataset Metadata & GDPR	GOVERN-1.2: Govern training data; ensure quality and representativeness	Annex IV §2(c): Training data description, provenance, and relevance	Clause 7.2.2: Data governance for AI	Fairness: Ensure training data is representative and unbiased
Section 4.8: Digital Signatures & Accountability	GOVERN-3.1: Define roles and responsibilities; ensure accountability	Art. 14: Human oversight measures; Annex IV §2(h): Audit trail	Clause 5.1: Leadership accountability for AI management system	Accountability: Ensure accountability throughout AI lifecycle
Section 4.9: Output	MEASURE-3.2: Monitor AI system	Art. 72: Post-market	Clause 9.2: Monitoring and	Robustness & Safety: Ensure



AI Supplement Requirement	NIST AI RMF 1.0	EU AI Act (Reg 2024/1689)	ISO/IEC 42001:2023	OECD AI Principles
Monitoring & Drift Detection	performance; detect drift and anomalies	monitoring by providers; Annex VII : Obligations	measurement of AI system performance	reliable and safe operation over time
Section 4.10.2: Human Oversight	MANAGE-4.1 : Incorporate human oversight in AI system operation	Art. 14 : Human oversight measures for high-risk AI	Clause 7.3.1 : Human oversight arrangements	Human-centered values : Design AI systems with humans in control
Section 4.10.3: Fairness & Bias Mitigation	MEASURE-2.10 : Assess and document AI system fairness and bias	Art. 10(2)(f) : Bias monitoring and mitigation measures	Clause 6.1.3 : Address bias and fairness risks	Fairness : Inclusive growth, sustainable development, well-being
Section 4.10.4: Explainability	MEASURE-2.3 : Ensure AI system outputs are interpretable and explainable	Art. 13(3)(b) : Information to users; explanation of decision logic	Clause 7.2.4 : Explainability of AI system decisions	Transparency : Enable understanding of AI decisions
Section 5: Assessment Methodology	GOVERN-4.1 : Conduct regular AI system assessments and audits	Art. 43 : Conformity assessment procedures for high-risk AI	Clause 9.3 : Internal audit of AI management system	Accountability : Independent evaluation and verification

Key Observations:

1. **Strong Alignment with EU AI Act**: Most requirements (Sections 4.2–4.9) directly map to EU AI Act Annex IV technical documentation requirements for high-risk AI. Organizations certified under this supplement will have substantial evidence for EU AI Act conformity assessment.
2. **Comprehensive Coverage of NIST AI RMF**: Requirements span all four RMF functions: GOVERN (accountability, governance), MAP (context, risk), MEASURE (validation, monitoring), MANAGE (incident response, human oversight). This supplement operationalizes NIST AI RMF principles.



GBA

BMM Artificial Intelligence Supplement

3. **ISO/IEC 42001 Compatibility:** Requirements align with ISO 42001 clauses on AI system documentation (7.2), data governance (7.2.2), validation (7.2.5), and monitoring (9.2). Organizations pursuing ISO 42001 certification can leverage BMM AI assessment artifacts.
4. **OECD Principles Embodiment:** All five OECD AI Principles (Transparency, Accountability, Fairness, Robustness & Safety, Human-centered values) are operationalized through specific technical requirements. This grounds abstract principles in auditable practice.

Strategic Value:

Organizations certified under this supplement can demonstrate compliance with multiple frameworks simultaneously, reducing audit burden and accelerating market access in multiple jurisdictions (EU, US federal procurement, international markets).



Appendix C: [INFORMATIVE] Quantum Readiness Overview

Purpose

Quantum computing may eventually compromise widely used public-key cryptography. This appendix provides the limited quantum context needed by BMM AI Supplement assessors. It is informative and does not create separate certification requirements. The operative requirements remain the signature, key-management, provenance, monitoring, and incident-response controls in Section 4. A more detailed treatment of the threats and opportunities pertaining to quantum computers can be found in the BMM Quantum Computing Supplement [6].

Why it matters

A sufficiently capable quantum computer could break RSA, ECDSA, Diffie-Hellman, and elliptic-curve Diffie-Hellman. These mechanisms are commonly used for signing, key exchange, and authentication in blockchain-enabled AI systems. This matters because evidence records, signatures, and encrypted communications may need to remain trustworthy for many years. Organizations should therefore avoid assuming that a cryptographic choice suitable today will remain suitable throughout a system's retention or operating life.

Assessor focus

- Assess the controls already required by Section 4.8: approved signature algorithms, key registries, trusted timestamps, authorization matrices, key rotation, key revocation, and formal key-management infrastructure where applicable.
- For Type 3 systems, confirm that the organization has documented cryptographic agility: the ability to replace signing, key-exchange, or hash mechanisms without redesigning the entire solution.
- For Type 3 systems with long-lived audit evidence, confirm that cryptographic migration is considered in lifecycle planning, record-retention decisions, and incident response. This is evidence of prudent risk management, not a requirement to deploy a specific post-quantum algorithm under this supplement.
- Continue to treat SHA-256 hash commitments as acceptable for the artifact-integrity controls in Section 4, while recognizing that the organization should periodically review cryptographic choices as authoritative standards and threat conditions change.



GBA

BMM Artificial Intelligence Supplement

Further guidance

Detailed post-quantum cryptography algorithm selection, migration planning, and quantum-specific threat analysis are outside the scope of this supplement. Assessors and solution providers should consult current NIST post-quantum cryptography guidance and applicable sector, national-security, and procurement requirements.



GBA

BMM Artificial Intelligence Supplement

Bibliography

The following references support the frameworks and guidance cited in this supplement.

1. “Inclusion of context awareness in machine learning in order to boost resiliency” (<https://www.mdpi.com/1999-4893/16/3/165>) by the Department of Computer Science, Sumy State University and the Department of Computer Systems, Networks and Cybersecurity, National Aerospace University “KhAI”, both from Ukraine
2. NIST AI RMF 1.0 <https://www.nist.gov/itl/ai-risk-management-framework>
3. EU AI Act (Reg. 2024/1689) <http://data.europa.eu/eli/reg/2024/1689/oj>
4. ISO/IEC 42001:2023 <https://www.iso.org/standard/42001>
5. OECD AI Principles (2024) <https://www.oecd.org/en/topics/sub-issues/ai-principles.html>
6. BMM Quantum Computing Supplement [In progress; link not yet available]
7. Blockchain Maturity Model v2.0 <https://gbaglobal.org/blockchain-maturity-model/>