Model AI Legislation Framework

Tier 2 – Technical Basis

Version 2.2 | December, 2025

A technical reference for risk assessment, scoring methodology, and implementation structure

This tier provides the technical foundation supporting the Model AI Legislation Framework, detailing the analytical methods, scoring logic, and risk evaluation structures referenced in **Tier 1.**

Developed by AI Safety International

Preface for Legislators and Legislative Staff

This document is intended to support legislators, legislative staff, regulators, and reviewers who are developing or evaluating artificial intelligence policy using the **Model AI Legislation Framework**.

Tier 1 of this framework establishes a **risk-based legislative foundation** without prescribing specific technologies, architectures, or policy outcomes. Tier 2 serves as a **technical companion** to that foundation. It explains the analytical methods referenced in Tier 1 so they may be understood, evaluated, and applied consistently during legislative development, regulatory review, or implementation planning.

This tier is **not proposed statutory language**. It does not instruct lawmakers on how to draft bill text, nor does it mandate specific regulatory approaches. Instead, it provides a shared technical reference that legislative teams may rely upon when determining:

- what types of risks require consideration,
- how those risks may be evaluated in a structured and repeatable manner, and
- how documentation and oversight requirements can be framed without embedding technical detail directly into statute.

Legislative drafters typically incorporate concepts from this tier indirectly, by:

- referencing risk assessment or documentation requirements in statute,
- delegating technical specificity to agencies or rulemaking bodies, or
- citing recognized analytical methods as acceptable means of compliance.

This approach allows legislation to remain durable over time while still being grounded in measurable, reviewable processes.

Readers should understand Tier 2 as **enabling material**. It supports informed decision-making by explaining how risk-based evaluation functions in practice, without requiring adoption of any single methodology. Jurisdictions or organizations may adopt Tier 1 independently, adopt Tier 2 as a supporting reference, or substitute alternative technical standards consistent with the principles established in the framework.

Together, Tier 1 and Tier 2 are designed to reduce ambiguity, avoid over-specification, and provide legislators and their staff with a clear analytical foundation upon which enforceable and adaptable policy may be built.

1.0 Purpose and Scope of the Technical Basis

1.1 Purpose

This document provides the technical foundation supporting the **Model AI Legislation Framework**. It describes the analytical methods, scoring logic, and evaluation structures referenced in **Tier 1 — Foundational Framework**, with the goal of enabling consistent, risk-based assessment of artificial intelligence systems.

The purpose of this tier is to:

- Explain how identifiable risks may be evaluated using structured analysis
- Provide technical clarity without prescribing regulatory outcomes
- Support oversight, review, and implementation across jurisdictions and sectors

This document is intended to function as a **technical companion**, not as statutory language or regulatory mandate.

1.2. Scope

The Technical Basis applies to artificial intelligence systems that fall within the scope defined in Tier 1. It is designed to be:

- **Informative**, rather than prescriptive
- **Technology-neutral**, independent of specific model architectures
- **Adaptable**, allowing use by legislatures, agencies, organizations, or reviewers

Nothing in this tier alters the applicability thresholds, exclusions, or legislative intent established in Tier 1.

2.0 Relationship to Tier 1 (Foundational Framework)

2.1 Supporting Role

Tier 2 exists solely to support the Foundational Framework. It provides technical detail for concepts that are intentionally stated at a higher level in Tier 1, including:

- Risk-based evaluation
- Failure mode identification
- Proportional mitigation
- Documentation and review practices

Tier 1 establishes *what* must be considered.
Tier 2 explains *how* such considerations may be carried out.

2.2 Independence of Tier 1

The Model AI Legislation Framework does not require adoption of this Technical Basis to remain valid. Jurisdictions or organizations may:

- Adopt Tier 1 without Tier 2
- Substitute alternative technical methodologies
- Use Tier 2 selectively as reference material

This separation is intentional and preserves legislative flexibility.

2.3 Non-Modification of Legislative Intent

Nothing in this tier:

- Expands regulatory authority
- Introduces enforcement mechanisms
- Alters definitions, scope, or exclusions established in Tier 1

Tier 2 does not create obligations beyond those explicitly defined elsewhere.

3.0 Risk-Based Assessment Methodology Overview

3.1 Rationale for Structured Risk Analysis

Artificial intelligence systems that interact with humans may produce harm through a range of failure modes that are:

- context-dependent
- probabilistic rather than deterministic
- emergent at scale

Unstructured or purely qualitative assessments are insufficient to address these characteristics consistently. Structured risk analysis methods provide:

- repeatability
- traceability
- documentation suitable for oversight and review

For this reason, Tier 1 references risk-based evaluation rather than categorical or outcome-based regulation.

3.2 Transferability of Engineering Risk Models

Risk assessment methodologies developed in engineering disciplines are applicable to artificial intelligence systems when adapted appropriately. Such methods:

- do not require assumptions about intent
- focus on observable failure and effect
- allow proportional response rather than binary judgment

This framework draws conceptually from established engineering practices without mandating their exclusive use.

3.3 Accepted Risk Analysis Models

One example of an accepted structured methodology is **Failure Mode and Effects Analysis (FMEA)**, adapted for artificial intelligence contexts. Equivalent methodologies may also be used, provided they demonstrate:

- systematic identification of failure modes
- defined criteria for assessing impact and likelihood
- documentation sufficient for independent review

The choice of specific analytical tools remains at the discretion of the implementing body.

4.0 Failure Mode Identification

4.1 Definition of Failure Mode in an Artificial Intelligence Context

For the purposes of this framework, a *failure mode* is a specific, identifiable way in which an artificial intelligence system may produce harm, unintended consequence, or material risk to users or society when operating as designed or when interacting with real-world conditions.

In the context of artificial intelligence, failure modes may arise from:

- system behavior rather than component malfunction,
- probabilistic outputs rather than deterministic error,
- interaction effects between system design, deployment context, and user behavior.

Failure mode identification focuses on **observable outcomes and effects**, not on inferred intent or internal model mechanics.

4.2 Categories of AI-Related Failure Modes

Failure modes may be grouped into non-exclusive categories to support systematic analysis. Common categories include, but are not limited to:

- **Informational failures**
  Outputs that are materially false, misleading, or presented with unwarranted confidence.
- **Decision-influence failures**
  Outputs that improperly shape, distort, or over-weight user judgment or decision-making.
- **Emotional or psychological impact failures**
  Outputs that contribute to emotional dependency, distress, manipulation, or erosion of user agency.
- **Contextual misalignment failures**
  Outputs that are inappropriate given the user's situation, vulnerability, or stated intent.
- **Scale and repetition failures**
  Harms that emerge primarily through repeated interaction, broad deployment, or cumulative exposure rather than single outputs.

These categories are illustrative and do not limit identification of additional failure modes relevant to specific systems or contexts.

4.3 Sources of Failure Modes

Failure modes may originate from multiple sources, including:

- **System design characteristics**
  Training data composition, optimization objectives, or output constraints.
- **Deployment context**
  Intended use, user population, and environment of operation.
- **Human-system interaction patterns**
  User reliance, misuse, misunderstanding, or repeated engagement.
- **Operational scale**
  Volume of interactions, geographic reach, or integration into critical processes.

Effective failure mode identification considers these sources collectively rather than in isolation.

4.4 Identification Process

Failure mode identification should be conducted using a structured, documented process. Such a process typically includes:

- Description of the AI system and its intended use
- Identification of plausible adverse outcomes associated with system outputs or behaviors
- Documentation of conditions under which each failure mode may occur
- Identification of affected users or stakeholder groups

The objective is not to enumerate all conceivable risks, but to identify **plausible, material failure modes** that warrant evaluation under the risk-based framework.

4.5 Documentation of Identified Failure Modes

Each identified failure mode should be documented with sufficient clarity to support subsequent risk evaluation and review. Documentation typically includes:

- a concise description of the failure mode,
- the nature of the potential harm or effect,
- the context in which the failure may arise,
- the user or system impact area affected.

Clear documentation supports traceability, enables independent review, and allows reassessment as systems or deployment conditions evolve.

Legislative Context Note (Informative)

In legislative or regulatory contexts, failure mode identification is commonly reflected through **assessment or documentation requirements**, rather than through enumeration of specific risks in statute. This approach allows laws to remain adaptable while ensuring that material risks are systematically considered.

5.0 Severity, Occurrence, and Detection Scales

Risk evaluation within the Model AI Legislation Framework relies on three complementary dimensions: **Severity**, **Occurrence**, and **Detection**. Together, these dimensions support structured comparison of identified failure modes and enable proportional risk management without reliance on categorical judgments.

The use of defined scales promotes consistency, transparency, and reviewability across assessments.

5.1 Severity (S)

5.1.1 Definition

Severity represents the magnitude of impact on users or society should a failure mode occur and reach the affected party.

Severity assessment focuses on **effect**, not intent. It evaluates harm based on outcomes rather than system design objectives or developer motivation.

5.1.2 Severity Considerations

Severity may encompass one or more of the following impact domains:

- emotional or psychological harm
- impairment of judgment or decision-making
- legal, financial, or reputational harm
- physical safety or wellbeing
- erosion of user autonomy or trust

Severity is assessed independently of how frequently a failure may occur.

5.1.3 Severity Scaling

Severity is typically expressed using an ordinal scale (for example, low to critical), with higher values representing greater potential harm. Scale definitions should be:

- clearly described,
- consistently applied,
- and documented with rationale.

Severity assessment should reflect **reasonable worst-case impact** within the anticipated deployment context.

5.2 Occurrence (O)

5.2.1 Definition

Occurrence represents the likelihood that a given failure mode will manifest under expected operating conditions.

Occurrence assessment evaluates probability, not impact.

5.2.2 Occurrence Considerations

Factors influencing occurrence may include:

- deployment scale and user volume
- frequency and duration of system use
- system safeguards or constraints
- environmental or contextual variability
- historical incident data, where available

Occurrence assessment should consider both individual interactions and cumulative exposure effects.

5.2.3 Occurrence Scaling

Occurrence is typically expressed using an ordinal scale representing relative likelihood. Scale values should be assigned based on:

- evidence where available,
- reasonable inference where data is limited,
- and documented assumptions.

Uncertainty should be acknowledged rather than obscured.

5.3 Detection (D)

5.3.1 Definition

Detection represents the likelihood that a failure mode will be identified **before** resulting harm occurs, or before harm escalates.

Detection assessment evaluates monitoring capability, not prevention.

5.3.2 Detection Considerations

Detection may depend on:

- system monitoring and logging mechanisms
- user reporting pathways
- escalation and intervention processes
- signal clarity and response timeliness

Lower detection capability corresponds to higher risk, even when severity or occurrence is moderate.

5.3.3 Detection Scaling

Detection is typically expressed using an ordinal scale where higher values represent **lower detectability**. Scale definitions should be explicit and consistently interpreted.

Detection assessment should reflect real-world operational conditions, not theoretical monitoring capacity.

5.4 Consistency and Documentation of Scoring

5.4.1 Scoring Discipline

Severity, Occurrence, and Detection scores should be:

- assigned independently,
- supported by written rationale,
- reviewed for internal consistency.

Scores are not intended to be precise measurements, but **structured judgments** supported by evidence and reasoning.

5.4.2 Avoidance of Mechanical Scoring

Risk evaluation should not rely solely on numerical calculation. Scores inform prioritization but do not replace human judgment, contextual awareness, or oversight.

Documentation of assumptions and uncertainties is essential to maintaining analytical integrity.

Legislative Context Note (Informative)

In legislative and regulatory settings, these scales are typically referenced as **elements of an assessment process**, rather than embedded directly in statute. This approach preserves flexibility while enabling consistent review and accountability.

6.0 Risk Priority Determination

Risk priority determination integrates **Severity (S)**, **Occurrence (O)**, and **Detection (D)** assessments to support comparative evaluation of identified failure modes. The purpose of this step is to enable **prioritization**, not to impose absolute thresholds or automated decisions.

Risk priority determination informs where attention, mitigation, and oversight resources should be directed.

6.1 Integrative Risk Evaluation

Risk priority is determined by considering the combined effect of:

- the magnitude of potential harm (Severity),
- the likelihood of manifestation (Occurrence),
- and the likelihood of early identification or intervention (Detection).

These dimensions are evaluated together to establish **relative risk posture** among identified failure modes.

No single dimension is sufficient on its own to determine priority.

6.2 Relative Ranking Rather Than Absolute Thresholds

Risk priority determination emphasizes **relative ranking**, not categorical pass/fail judgments.

Failure modes are typically assessed in relation to one another to identify:

- higher-priority risks requiring closer review or mitigation,
- moderate risks requiring monitoring or contextual controls,
- lower-priority risks that may be documented and revisited periodically.

This approach avoids rigid thresholds that may be inappropriate across different deployment contexts or jurisdictions.

6.3 Interpretive Use of Composite Risk Indicators

Composite risk indicators (including numerical combinations or visual representations) may be used to assist prioritization. Such indicators:

- summarize structured assessments,
- support comparison across failure modes,
- facilitate communication with reviewers and oversight bodies.

Composite indicators do not replace analytical judgment. They serve as **decision support**, not decision authority.

6.4 Context Sensitivity

Risk priority determination must account for context, including:

- system purpose and intended use,
- user population and vulnerability,
- deployment scale and duration,
- availability of mitigation or intervention mechanisms.

A failure mode with moderate inherent risk may warrant higher priority in certain contexts, while a higher-severity failure may warrant lower priority if occurrence is extremely limited and detection is strong.

6.5 Documentation of Priority Decisions

Priority determinations should be documented with sufficient clarity to support review. Documentation typically includes:

- the basis for relative ranking,
- key assumptions influencing prioritization,
- rationale for escalation or de-prioritization decisions.

Clear documentation supports transparency, accountability, and reassessment as systems or conditions change.

6.6 Non-Mechanical Judgment

Risk priority determination is not intended to be a mechanical or automated process. Structured analysis supports judgment but does not eliminate the need for:

- professional evaluation,
- contextual reasoning,
- oversight and review.

Human judgment remains central to interpreting risk assessments and determining appropriate responses.

Legislative Context Note (Informative)

In legislative and regulatory frameworks, risk prioritization is commonly reflected through **graduated obligations**, **documentation requirements**, or **review triggers**, rather than fixed numerical thresholds. This preserves adaptability while maintaining accountability.

7.0 Mitigation and Control Strategies

Mitigation and control strategies address identified failure modes in proportion to their assessed risk priority. The purpose of mitigation within this framework is to **reduce the likelihood, severity, or impact of harm**, not to eliminate all risk.

Mitigation strategies are selected based on context, feasibility, and effectiveness, and should be documented as part of the overall risk assessment process.

7.1 Types of Mitigation Strategies

Mitigation strategies may take a variety of forms, depending on the nature of the failure mode and the system's deployment context. Common categories include, but are not limited to:

- **Design-level controls**
  Adjustments to system behavior, output constraints, or response logic intended to reduce harmful outcomes.
- **Operational controls**
  Monitoring, escalation, intervention, or human-in-the-loop mechanisms applied during system operation.
- **User-facing measures**
  Disclosures, warnings, usage boundaries, or interface design choices that reduce misunderstanding or misuse.
- **Procedural controls**
  Review processes, update requirements, or operational guidelines governing system use.

These categories are illustrative and may be combined or adapted as appropriate.

7.2 Proportionality of Mitigation

Mitigation strategies should be **proportional to risk priority**.

Higher-priority risks typically warrant:

- more robust controls,
- closer monitoring,
- clearer documentation and review.

Lower-priority risks may warrant:

- documentation without immediate intervention,
- periodic reassessment,
- contextual or situational controls rather than structural changes.

Proportionality avoids both under-response and over-constraint.

7.3 Selection and Evaluation of Mitigation Measures

Mitigation measures should be evaluated based on:

- their expected effectiveness,
- feasibility within the deployment context,
- potential unintended consequences,
- impact on system functionality and user experience.

Selection should be supported by documented rationale rather than assumed effectiveness.

7.4 Residual Risk

Even after mitigation measures are applied, **residual risk** may remain. Residual risk reflects the level of risk that persists after reasonable controls have been implemented.

Residual risk should be:

- explicitly acknowledged,
- documented with rationale,
- reviewed periodically.

Acceptance of residual risk does not imply disregard for safety; it reflects recognition that zero risk is not achievable in complex systems.

7.5 Reassessment Following Mitigation

Mitigation measures may alter Severity, Occurrence, or Detection characteristics of a failure mode. Where appropriate, reassessment should be conducted to:

- confirm the effectiveness of controls,

- identify new or modified failure modes,
- update priority rankings.

Reassessment supports continuous improvement without requiring constant redesign.

Legislative Context Note (Informative)

In legislative and regulatory settings, mitigation is commonly reflected through **process requirements** rather than prescribed controls. Laws may require documentation of mitigation efforts or demonstration of risk reduction without specifying how mitigation must be implemented.

8.0 Documentation and Review Practices

Documentation and review practices ensure that risk assessment, prioritization, and mitigation activities remain **transparent, traceable, and subject to oversight**. The purpose of documentation within this framework is to support accountability and reassessment, not to impose administrative burden.

Effective documentation enables informed review by internal governance bodies, regulators, or judicial authorities where applicable.

8.1 Core Documentation Elements

Documentation associated with risk-based assessment typically includes:

- identification of assessed AI systems and deployment context,
- documented failure modes and associated risk evaluations,
- prioritization rationale and mitigation decisions,
- description of applied controls and residual risk,
- dates of assessment and review.

Documentation should be sufficient to explain **what was considered, how conclusions were reached, and why decisions were made**.

8.2 Documentation Proportionality

Documentation requirements should be **proportional to risk priority**.

Higher-risk systems or failure modes generally warrant:

- more detailed documentation,
- clearer justification of mitigation choices,
- more frequent review.

Lower-risk systems may be documented at a summary level, provided rationale and reassessment triggers are clearly stated.

This approach avoids unnecessary burden while preserving accountability.

8.3 Review Triggers

Risk assessments and associated documentation should be reviewed when material changes occur. Common review triggers include:

- significant system updates or retraining,
- changes in deployment context or user population,
- identification of new failure modes,
- credible incident reports or adverse outcomes.

Review triggers should be documented in advance to support consistency and predictability.

8.4 Internal and External Review Compatibility

Documentation practices under this framework are designed to support multiple forms of review, including:

- internal governance or ethics review,
- regulatory or agency oversight,
- independent audit or expert evaluation,
- judicial review where applicable.

Documentation should be understandable to reviewers without requiring access to proprietary model internals or trade secrets.

8.5 Record Retention and Accessibility

Records should be retained for a period sufficient to support review and accountability, consistent with applicable legal or organizational requirements.

Documentation should be accessible to authorized reviewers in a form that supports evaluation without exposing unnecessary or sensitive information.

Legislative Context Note (Informative)

Legislation implementing risk-based AI oversight commonly specifies **documentation and review obligations** rather than technical controls. This framework is designed to support such obligations by defining the types of records and review practices typically associated with responsible risk management.

9.0 Illustrative Application Examples (Non-Binding)

The examples in this section are provided solely to illustrate how the risk-based assessment methodology described in this tier may be applied across different classes of artificial intelligence systems. These examples are **non-binding**, **non-exhaustive**, and are not intended to prescribe specific controls, outcomes, or regulatory treatment.

The purpose of these examples is explanatory, not normative.

9.1 Conversational Artificial Intelligence Systems

Conversational AI systems interact directly with users through natural language exchanges and may influence understanding, judgment, or emotional state.

Illustrative failure modes in such systems may include:

- dissemination of materially misleading information,
- inappropriate responses to vulnerable user states,
- reinforcement of unhealthy dependency or reliance,
- failure to detect or respond appropriately to warning signals.

Risk assessment in conversational systems often emphasizes:

- contextual sensitivity,
- scale and repetition effects,
- detectability of harmful interaction patterns over time.

Mitigation may involve a combination of design constraints, monitoring mechanisms, and user-facing disclosures, selected in proportion to assessed risk.

9.2 Decision Support Systems

Decision support systems provide recommendations, analysis, or prioritization intended to inform human decision-making in areas such as finance, healthcare, employment, or public administration.

Illustrative failure modes may include:

- over-weighting of recommendations by users,
- opaque reasoning leading to uncritical reliance,
- propagation of biased or incomplete information,
- misalignment between system outputs and real-world conditions.

Risk assessment for decision support systems often considers:

- the significance of decisions influenced,

- the expertise and autonomy of intended users,
- availability of alternative information sources,
- consequences of incorrect or overconfident recommendations.

Mitigation strategies may focus on transparency, contextual framing, and safeguards that preserve human judgment.

9.3 Emotional or Behavioral Influence Systems

Some AI systems are designed to shape user behavior, engagement, or emotional response, either directly or indirectly.

Illustrative failure modes may include:

- manipulation of emotional state,
- erosion of user agency,
- reinforcement of harmful behavioral patterns,
- failure to recognize or respond to distress signals.

Risk assessment in such systems often emphasizes:

- vulnerability of the user population,
- duration and intensity of interaction,
- detectability of cumulative harm.

Mitigation approaches may include usage boundaries, escalation mechanisms, and periodic reassessment of interaction effects.

Interpretive Note

These examples demonstrate how the same risk-based methodology may be applied across different AI contexts without requiring distinct regulatory regimes for each system type. The underlying analytical approach remains consistent, while implementation details vary based on context and risk priority.

Legislative Context Note (Informative)

In legislative drafting, illustrative examples are typically placed outside statutory text or included as explanatory materials. Their purpose is to clarify application without constraining future interpretation or technological evolution.

10.0 Limitations and Non-Binding Nature

This tier is provided as **informative reference material** in support of the Model AI Legislation Framework. It is not statutory language, regulatory mandate, or enforcement guidance.

Nothing in this document:

- prescribes specific technologies, system architectures, or design choices,
- mandates particular mitigation measures or analytical tools,
- establishes legal thresholds, prohibitions, or compliance determinations.

10.1 Informative Role

Tier 2 explains technical concepts referenced in Tier 1 to support informed legislative, regulatory, and organizational decision-making. It does not require adoption of any specific methodology and does not restrict the use of alternative, equivalent risk assessment approaches.

10.2 Jurisdictional Flexibility

Jurisdictions, agencies, and organizations may:

- adopt Tier 1 independently of this technical basis,
- reference Tier 2 selectively,
- substitute alternative technical standards consistent with the principles established in the framework.

This flexibility is intentional and preserves adaptability across legal systems and regulatory environments.

10.3 No Expansion of Authority

This tier does not:

- expand regulatory authority,
- create new oversight obligations,
- or alter the scope or intent of Tier 1.

Any obligations arising from implementation of the framework derive solely from adopted legislation or regulatory action, not from this document.

10.4 Evolution and Revision

As artificial intelligence technologies and deployment practices evolve, technical methods may require refinement. Tier 2 is designed to support periodic review and update without requiring changes to foundational legislative language.

11.0 Transition to Tier 3

This tier has described the technical concepts and analytical methods referenced in the **Model AI Legislation Framework**, including failure mode identification, risk evaluation, prioritization, mitigation, and documentation practices.

Tier 3 — Adoption & Implementation Guidance builds upon this technical basis by addressing how the framework may be incorporated into:

- legislative drafting,
- agency rulemaking,
- organizational governance and review processes.

Tier 3 focuses on **practical pathways for adoption** without altering the principles, scope, or technical foundations established in Tiers 1 and 2.

Printed or downloaded copies may not reflect the most current revision. The authoritative version is maintained at aisafetyinternational.com.