BRIGHTBRIDGE LEARNING INTERNATIONAL

AI Risk Classification Standard

Consistent criteria for Low, Moderate, High, and Prohibited AI use

International Children's Education Nonprofit | Simulated AI Governance Consulting Engagement

Portfolio note

All organizational details, AI systems, roles, risks, scores, and data in this portfolio artifact are fictional or synthetic and are provided solely to demonstrate applied AI governance consulting methodology.

Version 1.0 | August 2026

1. Purpose

This standard provides a consistent, auditable method for determining the governance intensity required for each BrightBridge AI use case. Classification occurs before pilot or deployment and is repeated after material changes.

2. Risk Tiers

Tier	Definition	Minimum Governance
LOW	Limited potential to materially affect people; primarily internal drafting or administrative assistance using non-sensitive data.	Inventory; acceptable-use controls; owner; annual review.
MODERATE	Meaningful privacy, operational, reputational, or indirect stakeholder risk; human review remains central.	Inventory; risk classification; privacy/security/vendor review as triggered; documented controls; annual review.
HIGH	AI directly interacts with children or materially affects education, safeguarding, access, employment, sensitive data, or significant resource decisions.	Full assessment; safeguarding/privacy/security/fairness reviews; documented human oversight; control evidence; formal approval; ongoing monitoring.
PROHIBITED	Use conflicts with BrightBridge policy or creates risk that cannot be accepted or meaningfully controlled.	Do not pilot, procure, or deploy. Escalate any discovered use for immediate containment.

3. Scoring Methodology

Each factor is scored using the stated scale, then converted to a weighted contribution. The maximum score is 100. BrightBridge also applies automatic high-risk and prohibited-use overrides so that critical child or safeguarding risks cannot be diluted by lower scores in other categories.

Risk Factor	Weight	Scoring Guidance
Child interaction or child-affecting decision	20	0=None; 1=indirect; 2=limited; 3=material; 4=direct/consequential
Sensitive or high-impact data	15	0=public/non-sensitive; 1=internal; 2=personal; 3=sensitive; 4=highly sensitive child/safeguarding data
Safeguarding implications	20	0=none; 1=remote; 2=possible; 3=material; 4=direct safeguarding consequence
Decision impact and automation	15	0=drafting only; 1=informational; 2=advisory; 3=material recommendation; 4=automated/consequential
Bias, accessibility and inclusion exposure	10	0=low; 1=limited; 2=moderate; 3=high; 4=systemic or vulnerable-group impact
Cross-border data or jurisdiction complexity	5	0=single low-complexity region; 1=limited; 2=multi-country; 3=complex transfers/requirements
Manipulation, dependency or developmental vulnerability	10	0=none; 1=limited adult use; 2=indirect child impact; 3=child-facing persuasion/dependency; 4=heightened vulnerability
Vendor/model transparency limitations	5	0=full transparency/internal; 1=good; 2=partial; 3=limited/opaque

Score Bands

Score	Default Tier	Rule
0-24	Low	No high-risk override.
25-49	Moderate	No high-risk override.
50-100	High	Formal high-risk assessment and approval

Score	Default Tier	Rule
		required.
Any score	Prohibited	A prohibited-use rule overrides the numerical score.

Automatic High-Risk Overrides

- Direct child-facing conversational or recommendation AI.
- AI that materially influences access to educational support, safeguarding action, volunteer suitability, employment, or resource allocation.
- Processing of safeguarding narratives or highly sensitive child data.
- AI output used in a decision where an error could expose a child to material harm.
- AI that ranks, profiles, or infers behavior about children for consequential intervention.

BrightBridge Prohibited Uses

- Social scoring of children or families for generalized worthiness or character.
- Biometric or behavioral inference intended to infer sensitive traits unrelated to a narrowly defined safeguarding need.
- Manipulative AI designed to exploit children's developmental vulnerabilities or induce harmful dependency.
- Fully automated safeguarding decisions, case closure, or denial of essential educational support without meaningful human review.
- Covert AI surveillance of children or communities without a documented lawful, necessary, proportionate, and approved purpose.
- AI use that intentionally circumvents required human oversight, safeguarding review, or consent/notice requirements.

4. Mandatory Controls by Tier

Requirement	Low	Moderate	High	Prohibited
Inventory record	Yes	Yes	Yes	N/A - prohibited
Named owner	Yes	Yes	Yes	N/A
Privacy/security review	As triggered	Yes if personal data	Mandatory	N/A
Safeguarding review	If child-related	If child-related	Mandatory	N/A
Impact assessment	No	Targeted	Mandatory	N/A
Fairness/accessibility testing	As relevant	As relevant	Mandatory where differential impact plausible	N/A
Human oversight design	Basic	Documented	Formal with authority to override	N/A
Formal governance approval	Delegated	AI Governance Lead	AI Governance Committee	No approval permitted
Monitoring	Annual review	Defined metrics	KRIs + continuous/periodic monitoring	Containment
Independent assurance	No	Risk-based	Risk-based / selected high-risk systems	Incident review

5. Worked Classification Examples

System	Illustrative Score	Tier	Rationale
Student Dropout Risk Predictor	74	High	Child-affecting decision; educational intervention; sensitive socioeconomic proxies; cross-country use.
AI Learning Tutor	79	High	Direct child interaction; generative content; developmental vulnerability; learning dependence.

System	Illustrative Score	Tier	Rationale
Safeguarding Intake Assistant	88	High	Safeguarding narratives; highly sensitive data; false negatives may cause serious harm.
Translation & Localization AI	41	Moderate	Indirect child impact; human review for high-stakes content; moderate cultural/translation risk.
Donor Communications Assistant	18	Low	Internal drafting with human review; no child-facing decisions; limited sensitive data.
Candidate Matching Assistant	61	High	Employment ranking; fairness and discrimination exposure; material decision influence.

6. Classification Process

1. Confirm intended use, affected populations, system users, model type, data categories, automation level, countries, and vendor transparency.
2. Score each risk factor with documented evidence and rationale.
3. Apply automatic high-risk overrides.
4. Check BrightBridge prohibited-use rules.
5. Assign preliminary tier and required control set.
6. AI Governance Lead reviews Low/Moderate results; AI Governance Committee reviews High results.
7. Record tier in the inventory and link to required assessments and controls.

7. Reclassification Triggers

- New country or legal jurisdiction
- New child-facing functionality
- New sensitive data source
- Change from advisory to automated decision role
- New model or material vendor/model version
- Material change in training data or retrieval corpus
- New use case or affected stakeholder group
- Serious incident, fairness failure, or sustained performance drift
- Loss of vendor transparency or evidence

8. Classification Worksheet

Factor	Weight	Score	Evidence / Rationale
Child interaction or child-affecting decision	20	_____	Evidence / rationale: _____ _
Sensitive or high-impact data	15	_____	Evidence / rationale: _____ _
Safeguarding implications	20	_____	Evidence / rationale: _____ _
Decision impact and automation	15	_____	Evidence / rationale: _____ _
Bias, accessibility and inclusion exposure	10	_____	Evidence / rationale: _____ _
Cross-border data or jurisdiction complexity	5	_____	Evidence / rationale: _____ _

Factor	Weight	Score	Evidence / Rationale
Manipulation, dependency or developmental vulnerability	10	_____	Evidence / rationale: _____ _
Vendor/model transparency limitations	5	_____	Evidence / rationale: _____ _

Total weighted score: _____ / 100 Automatic High-Risk Override? Yes / No Prohibited Use? Yes / No Final Tier: _____

Portfolio deliverable result

The standard gives BrightBridge a repeatable way to triage every AI system and connect risk tier directly to mandatory governance, testing, approval, and monitoring requirements.