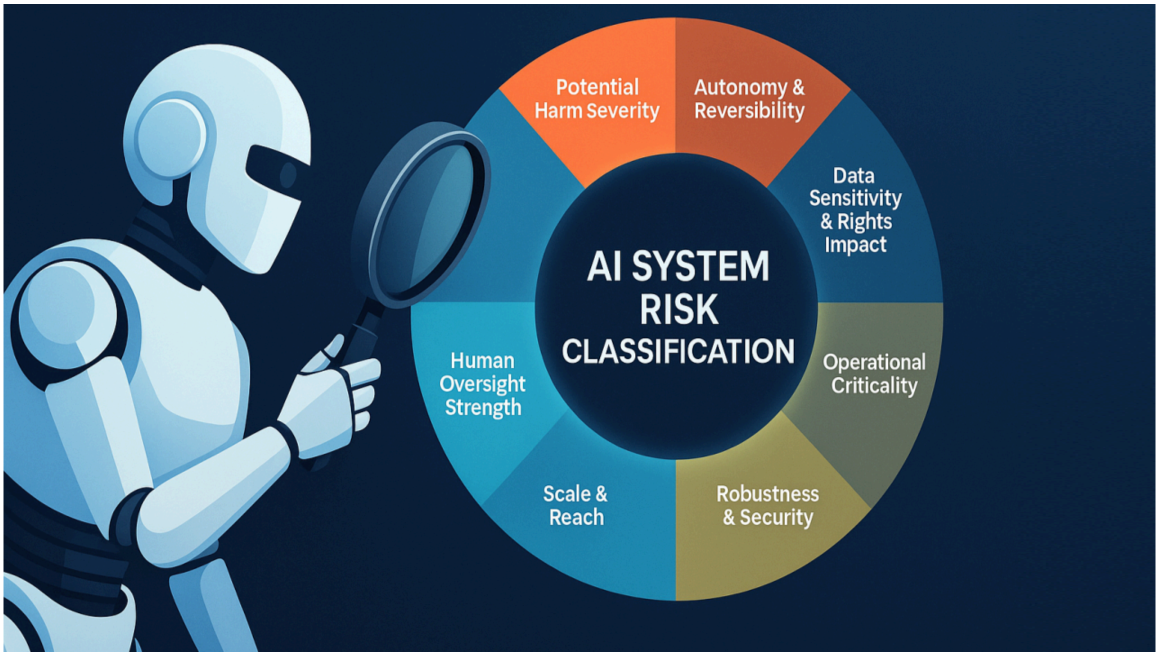


Image created with ChatGPT

Stop Guessing AI Risk: A Practical, Principle-Based Way to Classify AI Systems

Chris Fong
AI Governance | Technology-Driven Innovation & Transformation | Co-Founder@Fefifo (Exited)

November 18, 2025

If you've ever tried to classify an AI system by risk, you probably discovered something frustrating: **Everyone tells you that you *must* do it, but almost nobody tells you *how*.**

NIST, ISO 42001, and OECD all say "classify risks." The EU AI Act sorts systems into Prohibited, High-risk, Limited-risk, and Minimal-risk, but the categories depend almost entirely on predefined *use cases*.

Useful? Yes. Operational for enterprises? Not really.

Because real organisations aren't dealing with a neat list of "credit scoring," "biometric identification," or "toys with voice assistants." They're dealing with **hundreds of evolving AI systems — GenAI, classical ML, agentic workflows, SaaS tools with hidden AI features, and home-grown automations** that don't fit cleanly into any regulatory bucket.

That's why I developed a **first-principles, Responsible-AI-aligned, use-case-agnostic risk classification framework** accompanied by a control catalogue template. In this article, I'll explain **why organisations need a new approach and how this framework works.**

Why Existing Frameworks Aren't Enough

Most AI governance frameworks today fall into two camps:

1. Standards that tell you what to do, not how

NIST AI RMF, ISO 42001, OECD, etc. — they lay out the *expectations*:

- manage risks
- classify systems
- apply proportionate controls

But none of them provides a **workable scoring system** that enterprises can apply consistently across hundreds of use cases.

2. Regulations that classify by predefined use cases

The EU AI Act is groundbreaking. But its tiering works like this:

- Some use cases = automatically "high risk"
- Others = "limited risk" or "minimal risk"
- A few = banned

Great for regulators. Not enough for organisations dealing with emerging internal tools, vendor-provided copilots, or agentic systems that don't map to the law's categories.

> **"Use-case based classification solves compliance. Risk-factor based classification solves governance."**

Most organisations need both.

The Gap: We Need a Universal Way to Classify Any AI System

AI today is messy:

- Foundation models behave differently depending on prompts.
- SaaS tools embed AI without disclosing model architecture.
- Agentic systems chain multiple models, tools, and memory.
- Business teams adopt AI through Shadow IT, outside any governance process.

A use-case-based regulatory framework can't handle this complexity fully.

Organisations need a **consistent, repeatable, explainable way** to classify risks across all AI systems — regardless of vendor, architecture, or maturity.

Not because compliance demands it (it does).

But because **good governance depends on it.**

A First-Principles Approach: Classify Risks by What Really Matters

Instead of asking *"Which regulatory category does this belong to?"* we ask:

> “What risks does this system *actually* introduce into the organisation and to affected people?”

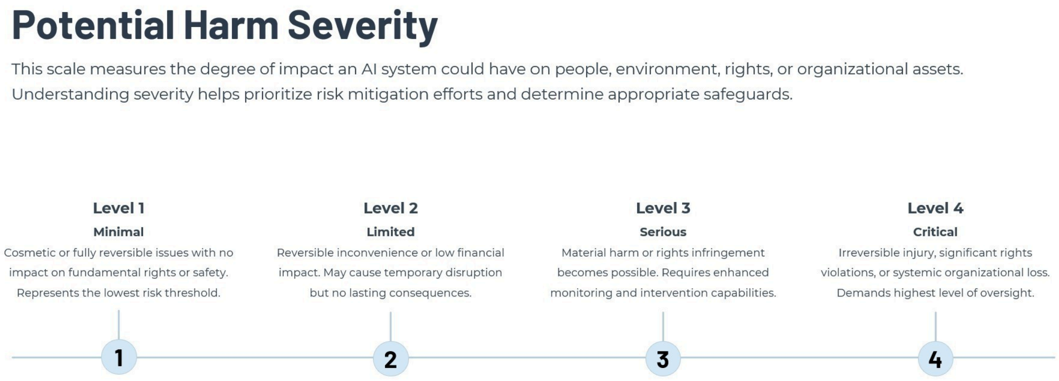
The framework evaluates seven core dimensions, each scored 1–4:

1. **Potential Harm Severity**
2. **Autonomy & Reversibility**
3. **Data Sensitivity & Rights Impact**
4. **Operational Criticality**
5. **Scale & Reach**
6. **Robustness & Cybersecurity Posture**
7. **Human Oversight Strength**

These seven factors are intentionally rooted in **Responsible AI principles** (fairness, accountability, safety, transparency, robustness, human agency).

Each factor has a crystal-clear definition.

No ambiguity. No philosophising. Just **operational clarity**.



Potential Harm Severity

This scale measures the degree of impact an AI system could have on people, environment, rights, or organizational assets. Understanding severity helps prioritize risk mitigation efforts and determine appropriate safeguards.

Level 1
Minimal
Cosmetic or fully reversible issues with no impact on fundamental rights or safety. Represents the lowest risk threshold.

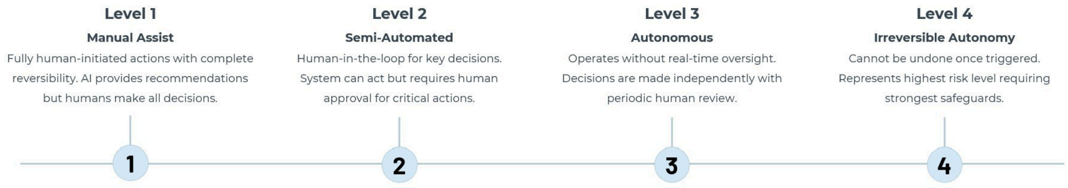
Level 2
Limited
Reversible inconvenience or low financial impact. May cause temporary disruption but no lasting consequences.

Level 3
Serious
Material harm or rights infringement becomes possible. Requires enhanced monitoring and intervention capabilities.

Level 4
Critical
Irreversible injury, significant rights violations, or systemic organizational loss. Demands highest level of oversight.

Autonomy & Reversibility

This dimension assesses the degree to which an AI system operates independently and whether decisions can be reversed or human intervention is possible. Higher autonomy with lower reversibility significantly increases risk exposure.



Level 1

Manual Assist

Fully human-initiated actions with complete reversibility. AI provides recommendations but humans make all decisions.

Level 2

Semi-Automated

Human-in-the-loop for key decisions. System can act but requires human approval for critical actions.

Level 3

Autonomous

Operates without real-time oversight. Decisions are made independently with periodic human review.

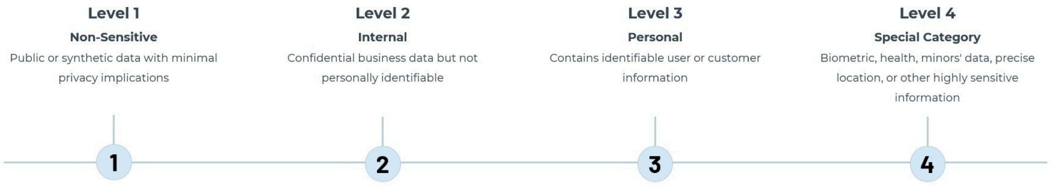
Level 4

Irreversible Autonomy

Cannot be undone once triggered. Represents highest risk level requiring strongest safeguards.

Data Sensitivity & Rights Impact

Evaluates the type and criticality of data the AI system processes. Special category data triggers enhanced legal obligations and requires additional protective measures under most privacy regulations.



Level 1

Non-Sensitive

Public or synthetic data with minimal privacy implications

Level 2

Internal

Confidential business data but not personally identifiable

Level 3

Personal

Contains identifiable user or customer information

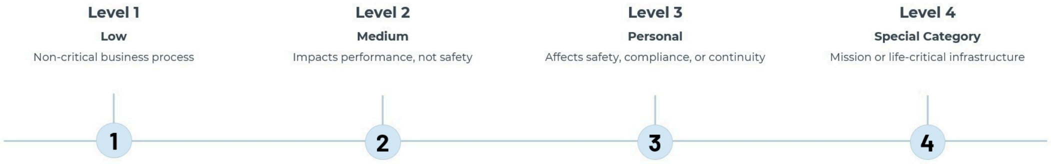
Level 4

Special Category

Biometric, health, minors' data, precise location, or other highly sensitive information

Operational Criticality

Evaluates the extent to which system failure disrupts essential operations.



Level 1

Low

Non-critical business process

Level 2

Medium

Impacts performance, not safety

Level 3

Personal

Affects safety, compliance, or continuity

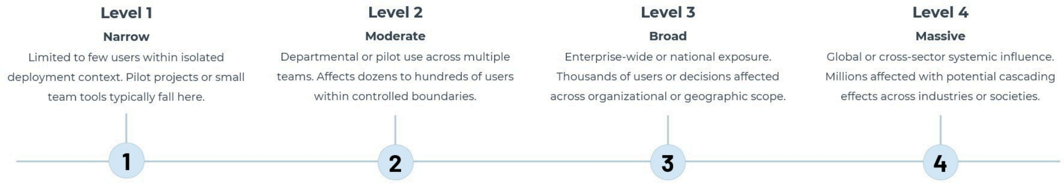
Level 4

Special Category

Mission or life-critical infrastructure

Scale & Reach

This factor assesses the breadth and frequency of the AI system's impact on people, processes, or decisions. Greater scale amplifies both potential benefits and risks, requiring proportional governance measures.



Level 1

Narrow

Limited to few users within isolated deployment context. Pilot projects or small team tools typically fall here.

Level 2

Moderate

Departmental or pilot use across multiple teams. Affects dozens to hundreds of users within controlled boundaries.

Level 3

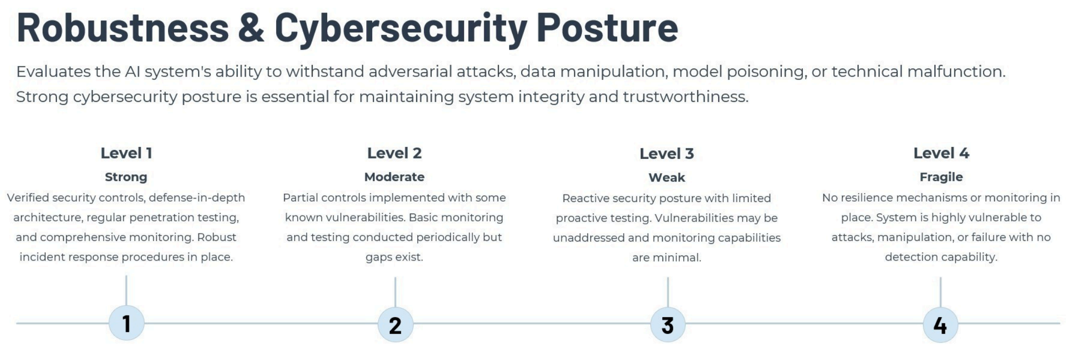
Broad

Enterprise-wide or national exposure. Thousands of users or decisions affected across organizational or geographic scope.

Level 4

Massive

Global or cross-sector systemic influence. Millions affected with potential cascading effects across industries or societies.



Robustness & Cybersecurity Posture

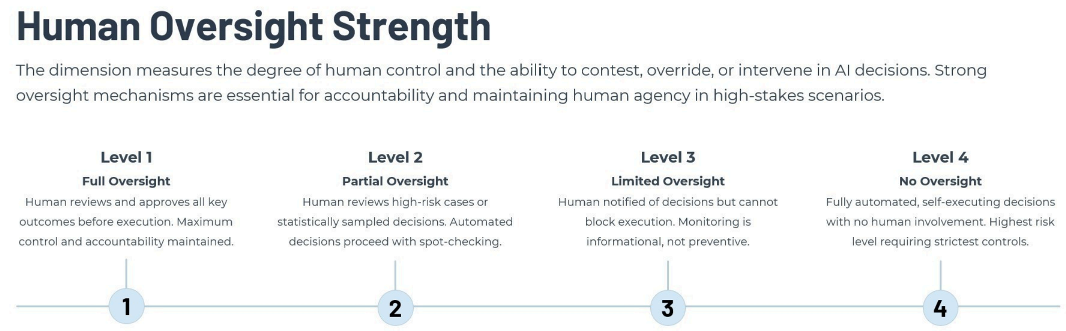
Evaluates the AI system's ability to withstand adversarial attacks, data manipulation, model poisoning, or technical malfunction. Strong cybersecurity posture is essential for maintaining system integrity and trustworthiness.

Level 1
Strong
Verified security controls, defense-in-depth architecture, regular penetration testing, and comprehensive monitoring. Robust incident response procedures in place.

Level 2
Moderate
Partial controls implemented with some known vulnerabilities. Basic monitoring and testing conducted periodically but gaps exist.

Level 3
Weak
Reactive security posture with limited proactive testing. Vulnerabilities may be unaddressed and monitoring capabilities are minimal.

Level 4
Fragile
No resilience mechanisms or monitoring in place. System is highly vulnerable to attacks, manipulation, or failure with no detection capability.



Human Oversight Strength

The dimension measures the degree of human control and the ability to contest, override, or intervene in AI decisions. Strong oversight mechanisms are essential for accountability and maintaining human agency in high-stakes scenarios.

Level 1
Full Oversight
Human reviews and approves all key outcomes before execution. Maximum control and accountability maintained.

Level 2
Partial Oversight
Human reviews high-risk cases or statistically sampled decisions. Automated decisions proceed with spot-checking.

Level 3
Limited Oversight
Human notified of decisions but cannot block execution. Monitoring is informational, not preventive.

Level 4
No Oversight
Fully automated, self-executing decisions with no human involvement. Highest risk level requiring strictest controls.

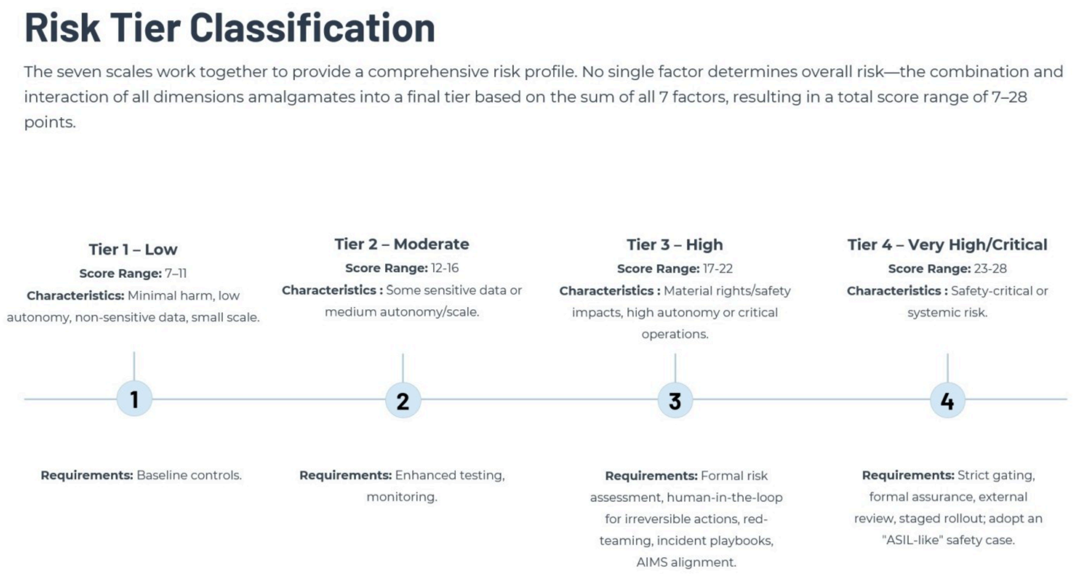
Define the Score → Determine the Tier → Apply the Controls

Once you score each factor, the total score (7–28) maps to four tiers:

- Tier 1 – Low (7–11)
- Tier 2 – Moderate (12–16)
- Tier 3 – High (17–22)
- Tier 4 – Very High / Critical (23–28)

What makes this powerful is that it sets the foundation for a **controls catalogue** to be mapped to each tier, flexibly tiering the rigor of your enterprises AI Risk Management Policies according to the risk criticality.

This creates **proportionate governance** — not overkill for low-risk tools, not under-governance for high-stakes systems.



Risk Tier Classification

The seven scales work together to provide a comprehensive risk profile. No single factor determines overall risk—the combination and interaction of all dimensions amalgamates into a final tier based on the sum of all 7 factors, resulting in a total score range of 7–28 points.

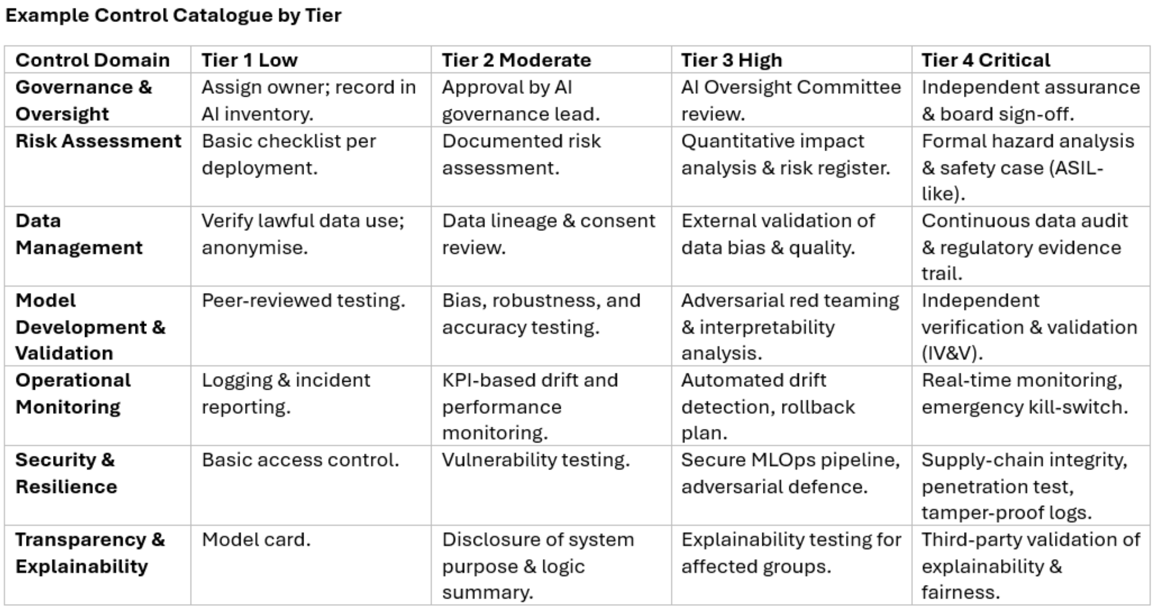
Tier 1 – Low
Score Range: 7–11
Characteristics: Minimal harm, low autonomy, non-sensitive data, small scale.
Requirements: Baseline controls.

Tier 2 – Moderate
Score Range: 12-16
Characteristics : Some sensitive data or medium autonomy/scale.
Requirements: Enhanced testing, monitoring.

Tier 3 – High
Score Range: 17-22
Characteristics : Material rights/safety impacts, high autonomy or critical operations.
Requirements: Formal risk assessment, human-in-the-loop for irreversible actions, red-teaming, incident playbooks, AIMS alignment.

Tier 4 – Very High/Critical
Score Range: 23-28
Characteristics : Safety-critical or systemic risk.
Requirements: Strict gating, formal assurance, external review, staged rollout; adopt an "ASIL-like" safety case.

Increasing controls rigor for higher risk systems



Example Control Catalogue by Tier

Control Domain	Tier 1 Low	Tier 2 Moderate	Tier 3 High	Tier 4 Critical
Governance & Oversight	Assign owner; record in AI inventory.	Approval by AI governance lead.	AI Oversight Committee review.	Independent assurance & board sign-off.
Risk Assessment	Basic checklist per deployment.	Documented risk assessment.	Quantitative impact analysis & risk register.	Formal hazard analysis & safety case (ASIL-like).
Data Management	Verify lawful data use; anonymise.	Data lineage & consent review.	External validation of data bias & quality.	Continuous data audit & regulatory evidence trail.
Model Development & Validation	Peer-reviewed testing.	Bias, robustness, and accuracy testing.	Adversarial red teaming & interpretability analysis.	Independent verification & validation (IV&V).
Operational Monitoring	Logging & incident reporting.	KPI-based drift and performance monitoring.	Automated drift detection, rollback plan.	Real-time monitoring, emergency kill-switch.
Security & Resilience	Basic access control.	Vulnerability testing.	Secure MLOps pipeline, adversarial defence.	Supply-chain integrity, penetration test, tamper-proof logs.
Transparency & Explainability	Model card.	Disclosure of system purpose & logic summary.	Explainability testing for affected groups.	Third-party validation of explainability & fairness.

Simple template to flexibly tailor to your organisation's AI risk management policy requirements

The Key Breakthrough: It Works for Any AI System

This classification method doesn’t care whether your system is:

- a retrieval-augmented chatbot
- a resume-screening model

- an agentic workflow automating system actions
- a fine-tuned LLM
- a vendor-provided SaaS copilot
- a predictive ML model from 2017
- an internal hackathon prototype
- a safety-critical decision engine

It applies equally to:

- internally-developed systems
- vendor-provided tools
- open-source or commercial models
- embedded AI features in software
- and even "non-AI" systems that quietly use AI internally

This universality is what organisations have been missing.

Why This Matters

AI systems are no longer "models." They're becoming **behaviours with consequences**.

And governance must match that shift.

A first-principles risk scoring framework gives an organisation:

- **Consistency** across all AI systems
- **Traceability** of decisions
- **Defensibility** in front of regulators
- **Proportionality** of controls
- **Clarity** for business teams
- **Actionability** for technical teams
- **Early detection** of high-risk initiatives

- **A shared language** across Legal, Risk, Cybersecurity, Data, and Engineering

Most importantly, it brings AI governance down from abstract principles to something practical that every organisation can actually operationalise.

Operationalising AI Governance Requires This Middle Layer

The industry is full of conceptual frameworks. Regulators are delivering top-down rules.

What’s missing is the **middle operational layer** — the part that tells organisations *how* to implement the principles and *how* to translate regulatory obligations into real-world practice.

AI system risk classification, built on Responsible AI principles, is one of the cornerstones of that operational layer.

It closes the gap between:

- **What standards require** and
- **What organisations actually need to do**.

A Final Thought

AI is moving faster than any governance structure ever has. The only way to keep up is to build frameworks that are:

- principled
- rigorous
- practical
- flexible
- future-proof

This AI system risk classification framework is one step in that direction.

If your organisation had to classify every AI system it uses or builds tomorrow — could it do so confidently?

If not, it's time to rethink your approach.

Chris Fong is a seasoned technology professional with deep experience spanning strategic planning, risk management, and transformation delivery across enterprise and technology ventures. He focuses on helping enterprises operationalise AI Governance with the goal to realise "Responsible, compliant, and effective use of AI across the organisation, enabling safe value creation while meeting regulatory and internal risk requirements".

Chris is a certified AI Governance Professional (AIGP) and AI Governance Architect (AIGA), Certified Information System Security Professional (CISSP), Certified Information System Auditor (CISA), AWS Certified AI Practitioner and IBM watsonx.governance: Technical Essentials trained.

Comments

89 · 28 comments · 11 reposts

Like Comment Share

Add a comment…

Most recent

Natasa Mihajlovic • 3rd+ 4d
Passionate about AI Governance, Quality, Compliance and Training.

Human Oversight Strength as described would not work in the clinical settings and life sciences industry (especially Level 3 or Level 4). So perhaps to consider to be specific in which industry that might work?

Like · 2 | Reply · 1 reply

Chris Fong Author (edited) 4d
AI Governance | Technology-Driven Innovation & Transfor...

Natasa Mihajlovic Thanks for the feedback, appreciate it! You're absolutely right, many use cases including those in clinical settings should not have limited / no oversight when deploying AI.

Any applicable regulatory mandate for strict human-in-the-lo ...more

Like | Reply

Paul Souhuwat • 3rd+ 4d •••
on the web

A great reminder Chris Entering the AI realm isn't about jumping into tools, as unfortunately some of of us still believe — it's about taking the right, low-risk steps first. Clear thinking, solid foundations, and honest risk classification matter more than flashy use cases. When we get the basics right, everything else becomes easier — and safer — to scale. It's just a

Like · 👍 1 | Reply · 1 reply

Chris Fong Author 3d •••
AI Governance | Technology-Driven Innovation & Transformation | ...

Paul Souhuwat So true, tools are critical for AI governance, but getting the first principles right behind what these tools are helping us with should be first priority, not just to implement what needs to be done, but more importantly to know when these tools lack. In the same tune, AI governance isn't just about compliance, it's more

Like | Reply

Sid Dutta in • 3rd+ 4d •••
Building the Future of Data Security in the age of AI!!

Governance, no matter what new form or shape you give it, doesn't reduce risks...just creates dashboards.

Like · 👍 1 | Reply · 2 replies

See previous replies

Sid Dutta in • 3rd+ 2d •••
Building the Future of Data Security in the age of AI!!

Chris Fong that's right...the missing link is "insights to action" as governance has been focussed primarily on the former.

Like | Reply

Jerushah Gracey MSc • 3rd+ 4d •••
Future Forward Founder x3: Adaptive Intelligence Layers™ | AI & Advance...

You gave a name to what so many teams are talking about now, the gap between principles and practice is widening as AI systems shift from "models" to behaviours with consequences.

The seven dimension framework you presented brings clarity the ...more

Like · 👍 1 | Reply · 1 reply

Chris Fong Author 3d •••
AI Governance | Technology-Driven Innovation & Transformation | ...

Jerushah Gracey MSc Exactly! The scoring framework and tiered controls catalogue template help to structure the assessment and dimensions for tiering of controls and governance according to your organisation's AI Risk Management Policy. But first, an operationalisable AI Risk Management policy has to be in place. I will

Like · 👍 1 | Reply

William P. in • 3rd+ 4d •••
Technology Executive | IT Strategy, Digital Transformation

Such a sharp articulation of the gap we've built libraries of principles and compliance checklists, but almost nothing in the middle that teams can actually run.

The standards say "classify the risk," but when you're dealing with c ...more

Like · 👍 2 | Reply

Sheriff O. in • 3rd+ 1w •••
Head of AI and Data Governance | Executive Director | Innovation | Data ...

Very insightful. This is a strong operational AI risk management approach combining the EU AI Act and other frameworks. Could you please share the template? Chris Fong

Like · 👍 2 | Reply

Rod Robertson • 3rd+ (edited) 5d •••
Chair at The Kitbag Group

Chris Fong. Hi Chris, Without the business ontology and the risk it mitigates, nothing will scale . I fear most organisations will have lists of use cases without them being integrated into the decisions and context/ risks behind them.

...more

Like · 👍 2 | Reply

Dr. Todd M. Price, MBA. ✓ • 3rd+ (edited) 5d •••
Author | Writer, Director, International Security Studies & Counte...

Threatindex.io and the DTM model has all frameworks and guardrails with ethcs in place. Where Operational meets academia! Global Counter-Terrorism Institute.

Home

Turn uncertainty into actionable intelligence. DTM's real-time risk scores help prevent conflicts & strengthen resilience for decision-makers worldwide

Like · 👍 2 | Reply

Danielle Hopkins • 3rd+ 4d •••
Seeking new opportunities leading AI Governance and Implementation a...

I hear this all the time, relating too having a fabulous suite of on paper governance. Literally zero in the way of practical implication and given the frameworks are now 'ancient' in AI development terms some are 3-4 years old and still evolving behind the times. Business is going to have to look at more practical implications that provide holistic and active governance

Like · 👍 1 | Reply

David Edwards in • 3rd+ 5d •••
Managing Director - Amment Risk Consulting

Chris Fong This is one of the most practical frameworks I've yet seen to codify, implement and embed the necessary governance of AI in all its forms - by calibrating risk in a standardised way that is capable of aggregation at the enterprise level - this is also a tool that can educate boards - without that education, effective AI governance won't get out of

Like · 👍 1 | Reply

Load more comments

Enjoyed this article?

Follow to never miss an update.

Chris Fong

AI Governance | Technology-Driven Innovation & Transformation | Co-Founder@Fefifo (Exited)

Follow

More articles for you

❮ ❯

This week's musings on tech contracts
The NIST Generative Artificial Intelligence Profile: a useful issue-spotter for contracts
TECH CONTRACTS ACADEMY

The NIST Generative Artificial Intelligence Profile: a useful issue-spotter for contracts

Tech Contracts Academy®

5 · 1 comment

EU AI Act - GPAI Code of Practice Published EU

Ankit Bhargava

112 · 5 comments

○ ○ ○ ○