Mastering AI Risk: Board and C-Suite Primer

Why traditional enterprise risk management breaks against generative AI, and what to build in its place: the Enterprise-Wide AI Risk Management (EW-AiRM™) framework, from first principles to implementation.

EW-AiRM™
by Human-Ai.Institute

INONEPARAGRAPH

Generative AI is probabilistic. It produces plausible outputs, not guaranteed ones, and its behaviour can change without anyone changing the code. Risk management built for deterministic systems cannot govern that on its own. **EW-AiRM™** is an open framework that augments, rather than replaces, existing enterprise risk management: **Three layers**, **six governance pillars**, an **evidence-based operational engine**, resilience against **eight AI Black Swan categories**, the **HAiPECR™** ethical overlay, and **Five Non-Negotiables** that no composite score can average away. This primer gives boards and executives the whole architecture in eight pages.

WHY AI BREAKS TRADITIONAL RISK MANAGEMENT

The decision to deploy generative AI is a threshold judgment, not an implementation detail. Deterministic software follows its code to specification; generative systems produce plausible outputs that are never guaranteed. **Hallucination** is a property of the architecture, not a failure of it. Behaviour can shift with input distributions or a provider-side **foundation model update**, while **shadow AI** and **device-embedded AI** spread the exposure well beyond the systems an organisation believes it operates.

BREAK 1

Deterministic assumptions

Legacy governance rests on one implicit promise: a validated system behaves the same way until someone changes it. Fixed thresholds, one-off validation reports and static segregation of duties all inherit that promise. Probabilistic systems break it; agentic systems, acting at scale, break it further.

BREAK 2

Static risk categories

Operational, legal, reputational and market buckets organise responsibility well, until a single AI incident cuts across all of them at once. Forcing a deepfake incident, an adversarial attack and a model reliability failure into the same legacy bucket obscures cause, dilutes remediation and erodes accountability.

BREAK 3

Temporal mismatch

Monthly reports and quarterly risk committees cannot govern exposure that moves in days: a foundation model update, a novel exploit technique, a silent shift in data distribution. Data drift is especially stealthy until its impact is obvious. Review cadences built for slower systems now leave blind intervals.

BREAK 4

The explainability gap

Internal representations are not human-readable: where deterministic systems allow root-cause tracing, generative models typically do not. Pre-deployment validation remains necessary but no longer sufficient; assurance must shift to output behaviour, calibration metrics and real-time anomaly detection.

THE THREE LINES OF DEFENCE, UNDER STRAIN BUT NOT OBSOLETE

The Three Lines remain the right accountability construct; each line simply needs augmentation as follows:
First line: continuous monitoring of model behaviour, not business outcomes alone.
Second line: a living risk taxonomy in place of a static quarterly register.
Third line: layered assurance that tests detection capability and response readiness, and feeds the central governance question: **should this system be deployed now, and under what restrictions?**

"The first AI governance question is not how to deploy. It is whether to deploy." — EW-AiRM™ Governance Maxim

This primer is an orientation to the framework; the full toolkit is published in the Wiley volume and at www.ewairm.com.

EW-AiRM™ at a glance

Augmentation, not replacement:
one architecture, three layers, an ethical overlay, and a floor that never moves.

EW-AiRM™
by Human-Ai.Institute

EW-AiRM™ is deliberately additive: it overlays AI-specific governance onto the ERM foundations an organisation already runs (ISO 31000, COSO ERM, the Orange Book, the IIA Three Lines Model etc.), so risk registers, committee charters and assurance cycles remain the backbone. Two strategic objectives anchor every decision the framework touches: **organisational survival** and **long-term sustainable benefit**.

THE THREE LAYERS

LAYER1·STRATEGIC

Six governance pillars

The decisions made before and around deployment: **six pillars**, from Strategic Alignment and Necessity Assessment (P1) through to **Through-the-Lifecycle Monitoring** and **Adaptability** (P6), each detailed on page 3.

LAYER 2 · OPERATIONAL

Grounded in evidence

Risk identification built on the MIT AI Risk Repository (MIT FutureTech): 7 risk domains and 24 subdomains drawn from **more than 1,700 categorised AI risks**, paired with the MIT AI Risk Mitigation Database of **831 evidence-based controls**.

LAYER3·RESILIENCE

Eight AI Black Swans

Low-probability, high-consequence events that risk registers miss, from emergent capabilities and **foundation-model concentration** through to **multi-agent emergence** and the **quantum cryptographic transition** (page 5).

HAIPECR™ — THE ETHICAL OVERLAY ACROSS ALL THREE LAYERS

HAiPECR, originating in 2023 as the **Human-Ai Paradigm for Ethics, Conduct and Risk**, runs across every layer as a **continuous ethical filter**; a documented pre-deployment assessment is itself Non-Negotiable NN2. Its seven dimensions map to the ten core principles of the UNESCO Recommendation on the Ethics of AI (2021), and it has been listed on the **OECD AI Policy Observatory's Tool catalogue** since April 2023.

- **H** **Human oversight and autonomy** · override real, not nominal
- **A** **Accountability** · a named individual, not a committee
- **i** **inclusivity and participation** · embedded, not a checkbox
- **P** **Protection of data, privacy and neural integrity**
- **E** **Ethics** · transparency, explainability, fairness: demonstrable
- **C** **Conduct risk** · AI conduct and human conduct, symmetric
- **R** **Rights and non-harm** · incl. intergenerational sustainability

THE FIVE NON-NEGOTIABLES — THE ALWAYS-ON FILTER

ANY ONE OF THESE AT A FAILING STATUS FORCES AN F-CRITICAL POSTURE, REGARDLESS OF THE COMPOSITE SCORE

NN1	**Named individual owner** for every deployed AI system, from deployment through decommissioning.
NN2	**HAiPECR pre-deployment assessment** for every new deployment.
NN3	**Tested, documented human override** capability, operable typically within four hours or less.
NN4	**Explicit, documented acceptance of residual AI risk** by an accountable person.
NN5	**AI incident reporting pathway with retaliation protection** (UK PIDA 1998; EU Directive 2019/1937).

WHAT EW-AIRM™ IS NOT

Not a replacement for ERM, cybersecurity, legal, compliance or model-development practice: it aligns with and supports them. Not a promise to eliminate rare, high-impact surprises, and not a substitute for executive judgement. The full toolkit lives in the Wiley volume and at www.ewairm.com.

The Strategic Layer: Six Pillars

The decisions made before and around deployment.
Each pillar carries a purpose, one core question, and the failure it is designed to prevent.

EW-AiRM™
by Human-Ai.Institute

The temptation to deploy AI simply because the technology exists is the most common and most expensive strategic failure. **The Strategic Layer forces the evaluation of value creation, risk exposure and organisational fit** before capability enthusiasm commits resources, and it turns that **evaluation into documented decisions: pause, go-ahead, or stage-gate.**

P1 · Strategic Alignment and Necessity Assessment

Purpose: test whether the initiative addresses a high-impact problem, what it costs if it fails, and whether it belongs on the roadmap at all. Customer insight gathered early is far cheaper than retrofitting a misfit programme later .

Core question: Should we deploy at all?

Typical failure: strategy misalignment; zombie projects kept alive by sunk-cost inertia because no go/no-go criteria were ever set.

P2 · Organisational Readiness

Purpose: establish whether people, processes, governance capacity and budget can actually carry the deployment, with an honest gap analysis and named personnel accountable for governance, assurance and incident response.

Core question: Are we ready, and is governance spend proportionate to quantified risk?

Typical failure: under-resourcing and misalignment; risk functions and business units operating in silos.

P3 · Data and Technological Maturity

Purpose: treat data readiness as the baseline precondition, not an afterthought: quality, provenance and lineage; auditable, authorised data pathways; automated checks for distributional drift, schema compliance and semantic consistency at ingress and along live inference paths.

Core question: Will our data and technology sustain reliable behaviour in production?

Typical failure: data quality and integration gaps surfacing after go-live, amplified by provider-side foundation model updates.

P4 · Risk Tolerance and Prioritisation

Purpose: define what the organisation will accept, contain and reject, with granular, tier-specific thresholds embedded in change control, incident playbooks and procurement terms, and reassessed after incidents, major model updates or shifts in use.

Core question: Where do we set risk appetite, and can we evidence the thresholds?

Typical failure: over-optimistic thresholds; appetite statements without triggers that convert observation into action.

P5 · Governance and Accountability

Purpose: place authority and liability for each live system on a named individual with real consequences: override authority when safety thresholds are breached, a single escalation channel, and explicit acceptance of residual risk as part of the role. Psychological safety is a practical control: reward those who pause a deployment for genuine risk.

Core question: Who owns each deployed AI system?

Typical failure: diffused accountability across committees; ownership as theatre without consequences.

P6 · Through-the-Lifecycle Monitoring and Adaptability

Purpose: continuous, tiered surveillance of living systems: scenario-based regression tests against curated benchmarks, statistical comparison of current outputs to historical distributions, and pre-specified triggers that automate governance review, with depth scaled to stakes.

Core question: How do we monitor and adapt as the system, the provider and the context change?

Typical failure: the gap between single-point sign-off and the next incident; drift invisible to ordinary business metrics.

"A risk appetite statement without specific thresholds is a press release, not a governance instrument."
— EW-AiRM™ Governance Maxim

The Operational Layer: The Engine

Where policy turns into practice:
an evidence-based taxonomy, a living risk register, and KXIs as the monitoring currency.

EW-AiRM™
by Human-Ai.Institute

THE MIT AI RISK REPOSITORY AS A LIVING TAXONOMY ENGINE

The Operational Layer activates the MIT AI Risk Repository (MIT FutureTech) as a working engine rather than a static reference file. Risks are organised by the harm they produce across seven domains: **Discrimination and Toxicity · Privacy and Security · Misinformation · Malicious Actors and Misuse · Human-Computer Interaction · Socioeconomic and Environmental · AI System Safety, Failures, and Limitations**. Each domain carries sub-domains backed by incident evidence and academic work, so a specific failure mode maps to a causal mechanism without inventing new categories.

7	24	1,700+	831	4
MIT RISK DOMAINS	SUBDOMAINS	CATEGORISED AI RISKS	EVIDENCE-BASED CONTROLS	CONTROL QUADRANTS

Once classified, each risk links to mitigations in the **MIT AI Risk Mitigation Database**: 831 evidence-based controls organised into four quadrants, **Q1 Governance and Oversight, Q2 Technical and Security, Q3 Operational Process, and Q4 Transparency and Accountability**. Controls attach to specific failure modes rather than hanging off systems as generic requirements.

THE ENHANCED AI RISK REGISTER

A register that updates, not a document that ages

New risks are classified into the seven domains as they emerge, with automated tagging, severity grading informed by historical incidents, and lifecycle tracking for each control from deployment to retirement.
Every control has a named owner, so responsibility is clear when a control fails or needs revision, and AI-specific risk signals flow directly into existing ERM reporting and escalation channels.

REGULATORY ALIGNMENT

Built to face supervisors

The framework is aligned with UNESCO, UNECE WP.6, NIST and ISO/IEC 42001 and maps to the EU AI Act.
It incorporates operational-resilience expectations from regimes such as PRA SS1/21, FCA PS21/3 and DORA, shifting the test from predicting every harm to staying within defined impact tolerances under severe disruption, and it draws on the OWASP Top 10 for LLM Applications for the technical security pillar.

KXIS: THE MONITORING CURRENCY

KXIs (Key X Indicators) are the umbrella currency for KPIs, KRIs and KCIs: an integrated dataset that exposes the dangerous mismatch where performance looks good while risk trends deteriorate. Cadence matches operational velocity: low-risk services may tolerate weekly checks, while high-risk services require near-real-time visibility with automated escalation. Every KXI carries a **named owner** and a **defined trigger-action mechanism**, converting observation into governance action rather than retrospective commentary. Human override rates are themselves tracked as control indicators, not dismissed as residual variance.

REPRESENTATIVE KXI

Model version currency

Flags any model version exceeding a predefined age, for example more than 60 days behind the latest security-relevant update, and initiates escalation or suspension automatically.

REPRESENTATIVE KXI

Data governance score

A composite measure of provenance, bias testing and consent coverage with a minimum acceptable threshold, say 80 out of 100, required before deployment proceeds.

"Accountability cannot live in a committee: A named human being must own every deployed AI system."
— EW-AiRM™ Governance Maxim

The Resilience Layer: Eight AI Black Swans

Preparing for what the other two layers cannot fully prevent:
low-probability, high- consequence events that risk registers miss.

EW-AiRM™
by Human-Ai.Institute

Traditional ERM assumes risks are visible, quantifiable and bounded by historical precedent. Generative systems introduce events that can arise without any record, materialise in weeks or days, and exceed the parameters of standard risk. The Resilience Layer therefore shifts the posture from anticipation to absorption: incident playbooks, cross- functional stress testing, escalation protocols and recovery arrangements, built around eight named categories.

1 Emergent capabilities
Behaviours absent during validation surface in production; small shifts in data or prompts reveal novel capability sets.

2 Systemic foundation model concentration
Dependence on a handful of upstream providers turns one provider-side change or outage into a market-wide event.

3 Alignment failures
The frontier risk: a model does exactly what it was trained to do, even where that diverges from human intent.

4 Multi-modal compound effects
Text, image, audio and code systems interact, producing compound failures no single-modality assessment predicted.

5 Adversarial breakthroughs
Novel attack techniques against models and pipelines that outpace controls designed for human-limited adversaries.

6 Supply chain compromises
Compromise of third-party components, data inputs or model provenance removes attestation and interrupts patching.

7 Multi-agent emergence
Autonomous agents interacting at scale produce collective behaviour, and failure modes, that no single agent exhibits.

8 Quantum cryptographic transition
Data harvested today decrypted tomorrow; without crypto-agility, signatures, audit trails and supply-chain attestations can collapse on short operational windows.

THREE RESILIENCE CAPABILITIES

CAPABILITY 1
Robust Foundations
Architecture, redundancy and contractual arrangements that keep unexpected behaviour from becoming operational collapse.

CAPABILITY2
Continuous Sensing
Detection tuned to interaction effects and slow-building hazards, catching issues before they cascade into the customer experience.

CAPABILITY 3
Adaptive Response
Rehearsed playbooks and escalation paths that let a named leader override automation and return the organisation to normal operation fast.

THE AGENTIC HORIZON

Many incidents labelled “user error” are better understood as collaborative misalignments between human cognitive bias and system design; reframing them keeps design teams accountable. Agentic systems that execute multi-step tasks with minimal human review will shrink today’s human checkpoint. Governance must prepare now: tighter intervention controls, richer observability, stronger acceptance criteria. **Firms that cannot keep pace will struggle to compete; firms that chase capability without restraint will struggle to survive.**

“The response to unknown unknowns is not anticipation, it is resilience.” — EW-AiRM™ Governance Maxim

Implementing EW-AiRM™ - at your scale

Three tiers, one architecture. Five phases, one direction of travel.
Governance intensity driven by actual exposure, not fear.

EW-AiRM™
by Human-Ai.Institute

SCALES BY TIER: SAME ARCHITECTURE, PROPORTIONATE DEPTH

CORE

Under 100 employees · 1 to 5 AI systems

Mostly COTS or API deployments. Minimum viable AI risk discipline: a lightweight system inventory, the mandatory HAiPECR pre-deployment assessment, basic incident logging, and a named individual accountable for each system from deployment through decommissioning.

STANDARD

100 to 1,000 employees · 5 to 25 systems

Full layer coverage at proportionate depth. Adds continuous monitoring, periodic model-performance reviews, a formal enhanced risk register, clearer operational procedures, and specialist governance roles created or shared across teams.

FULL

Over 1,000 employees · more than 25 systems

Comprehensive, living governance: full risk taxonomies, dedicated resilience protocols including Black Swan preparedness, and advanced control sets covering supply chain, model change management, agentic exposure and the quantum transition.

Whatever the tier, the Five Non-Negotiables remain the immutable floor. Tiers change the depth, never the architecture, and organisations progress between tiers as exposure grows, not as ambition does.

THE FIVE-PHASE ROADMAP

PHASE 1

Assess

Inventory every deployed and in-development system; map each against the six pillars; baseline-score the highest-risk systems and prioritise attention.

PHASE 2

Enhance

Educate the board and risk committee on the seven MIT domains; reclassify existing register entries against the taxonomy; close the quick, obvious gaps.

PHASE 3

Build

Construct the governance artefacts: pre-deployment assessment templates, control selections mapped to failure modes, escalation and override protocols.

PHASE 4

Implement

Deploy monitoring infrastructure and KXI dashboards for the highest-risk systems; set thresholds and escalation logic; train all three lines.

PHASE 5

Sustain

Continuous: track MIT taxonomy updates, horizon-scan for emerging risks, and make targeted future-proofing investments such as quantum-risk controls.

COMMON IMPLEMENTATION PITFALLS

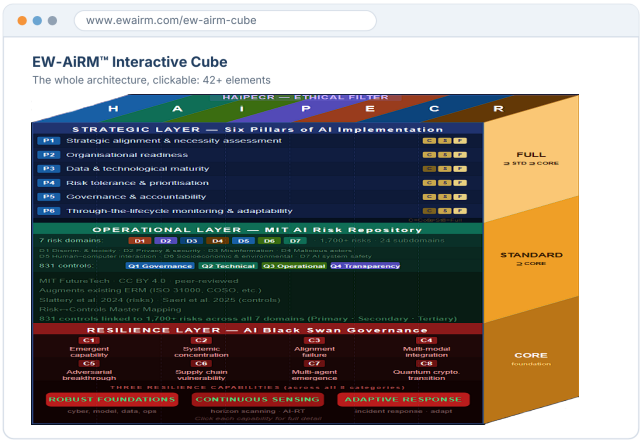
- **Governance and Calendar theatre.** Judging implementation by milestone dates rather than by whether governance actually functions. Maturity is measured in behaviour, not in project plans.
- **Fear-driven intensity.** Applying maximum controls everywhere. Risk-based allocation directs scarce governance capacity to the systems that drive consequential decisions; routine automations get a proportionate lighter touch.
- **Template replication.** Copying templates and risk registers verbatim. The published artefacts are for orientation and adaptation, not replication.
- **Cadence mismatch.** Monitoring rhythms that ignore how fast a system can change, especially where foundation-model providers push updates on their schedule, not yours.
- **Ownership drift.** Letting accountability migrate back into committees once the launch attention fades.

"Risk management, done well, does not slow innovation. It makes innovation sustainable."
— EW-AiRM™ Governance Maxim

The EW-AiRM™ Framework - in Software form

Free, browser-based instruments: one framework, seen from four different angles.
No sign-up, no install, and nothing leaves your browser (local install is also possible/available).

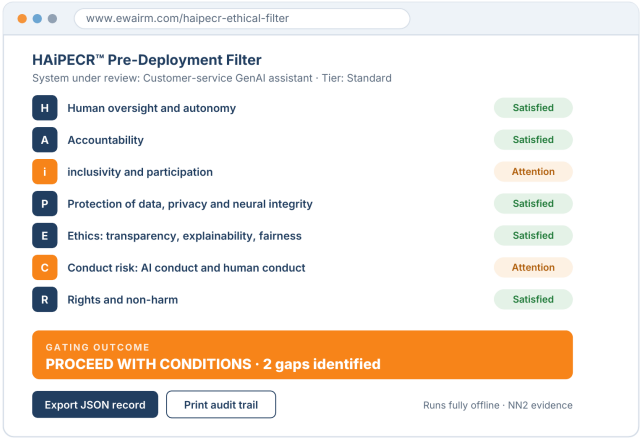
EW-AiRM™
by Human-Ai.Institute



EW-AiRM™ Interactive Cube

The whole architecture, clickable, in a COSO-style cube: three layers, six pillars, seven MIT domains, the HAiPECR™ overlay and all eight AI Black Swan categories, including the quantum cryptographic transition strip. The fastest visual orientation for a board.

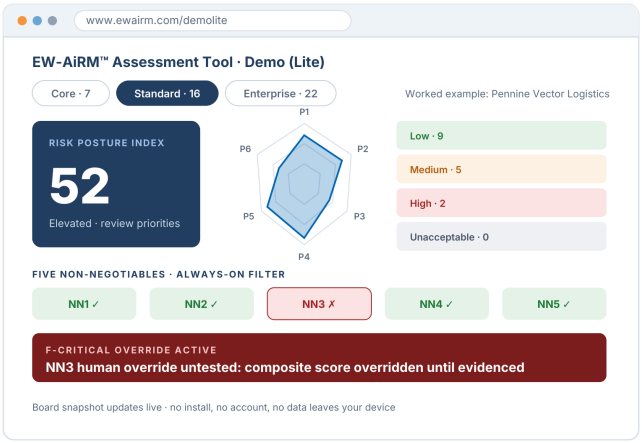
www.ewairm.com/ew-airm-cube



HAiPECR™ Pre-Deployment Ethical Filter

The seven-dimension ethical overlay as a deployment gating tool: one focused decision, taken before go-live and at material change reviews. Tool outputs: "Proceed", "Proceed with Conditions" or "Halt Deployment", with the specific gaps identified, and exports the documented record that NN 2 requires.

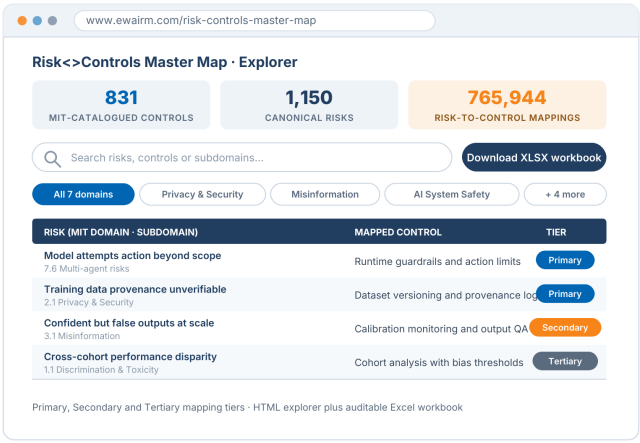
www.ewairm.com/haipecr-ethical-filter



EW-AiRM™ Assessment Tool · Demo (Lite)

Pick a tier, answer the questions, and watch a board-level snapshot update live: Risk Posture Index, risk counts, Non-Negotiables status and the F-Critical override that no composite score can average away. Includes the Pennine Vector Logistics worked example for teaching.

www.ewairm.com/demolite



Risk<>Controls Master Map

831 MIT-catalogued controls mapped to 1,150 canonical risks across the 7 domains and 24 subdomains, yielding 765,944 primary, secondary and tertiary mappings, with an interactive HTML explorer, an auditable Excel workbook and a methodology report.

www.ewairm.com/risk-controls-master-map

TRY THE FRAMEWORK, NOT JUST READ ABOUT IT

The tools and the book are built from the same canonical reference base: when the framework updates, they update together. Some tool previews above are illustrative renderings for print; run the live instruments at: **www.ewairm.com/companiontools**.

Prof. Markus Krebsz & the wider EW-AiRM™ ecosystem

Enterprise risk management's new stance, the person behind the framework, and where to go next.

Human-Ai.Institute

A NEW STANCE FOR ENTERPRISE RISK MANAGEMENT

Relying solely on deterministic risk methods in the age of generative AI is itself a serious vulnerability: it stalls innovation while leaving the organisation exposed. EW-AiRM™ offers the alternative stance: augment rather than replace, ask whether before how, and build resilience over prediction. **"Governance that looks right is not the same as governance that works."** The framework exists to close exactly that gap.

ABOUT THE AUTHOR

Prof. Markus Krebsz is the creator of EW-AiRM™ + HAiPECR™, and Founding Director of the Human-Ai.Institute and De-Risking Solutions Ltd. He served as AI Project Lead at UNECE Working Party 6, where he led the world's first Common Regulatory Arrangement (CRA) on products with embedded AI (ECE/TRADE/486, 2024), and contributed to the EU AI Office expert community and the General-Purpose AI Code of Practice process. He is a working-group member of IEEE P7007.1, a StandICT.eu Standards Maker, a ForHumanity Certified Auditor, Honorary Professor at the University of Stirling and Clinical Professor of Practice at Woxsen University, with > two decades of board-level risk experience across banking and regulation.

THE HUMAN-AI.INSTITUTE

An independent Think/Do Tank convening multilateral and multi-stakeholder dialogue on AI governance, engaging with the UN, UNESCO, the OECD, IEEE, the EU Commission and governments worldwide.

Home of EW-AiRM™ and a founding member of the UN University Global AI Network. **www.human-ai.institute**

FURTHER EW-AIRM™ RESOURCES

Companion tools suite

The Interactive Cube, HAiPECR Filter, Demo (Lite) and Risk<>Controls Master Map (page 7), all free and browser-based. **www.ewairm.com/companiontools**

HAiPECR on OECD.AI

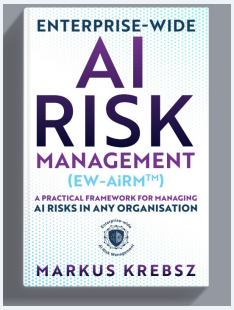
Listed on the OECD AI Policy Observatory's catalogue of tools and metrics since April 2023. **oecd.ai/en/catalogue/tools/haipecr**

Website, video and community

Framework, tools, insights and the Knowledge-Bot: **www.ewairm.com** · YouTube: **youtube.com/@EW-AiRM**

LinkedIn

linkedin.com/company/the-human-ai-institute and the thematic page **linkedin.com/company/enterprise-wide-ai-risk-management**



FORTHCOMING · WILEY FINANCE · 12 NOVEMBER 2026

Enterprise-Wide AI Risk Management (EW-AiRM™): A Practical Framework for Managing AI Risks in Any Organisation

The complete framework in one volume: the three layers, six pillars, HAiPECR™, the Five Non-Negotiables, the eight AI Black Swan categories and the implementation roadmap, with templates, KXI libraries, calibration matrices and worked examples throughout. ISBN 978-1-394-44647-6.
Pre-order: **www.amazon.co.uk/dp/1394446470**

CONTACT
contact@human-ai.institute
+44 (0) 79 85 065 045

WEB
www.ewairm.com
www.human-ai.institute

CONSULTING & TRAINING
Provided by De-Risking Solutions Ltd. and RiskAi.Ai