Chapter 1

What Are We Actually Talking About?

COPYRIGHTED MATERIAL

Ask 20 risk professionals to define artificial intelligence (AI) and you will get 20 different answers. I know this not from theory but from experience having asked this question in training rooms, at conferences, in board-level workshops, and in one-to-one conversations with some of the most experienced risk practitioners working today.

The answers range from the highly technical to the philosophical to the frankly baffled. They include 'machine learning' and 'the thing that writes emails' and 'ChatGPT' and 'something my teenagers use' and, from a chief risk officer at a major financial institution, the admirably honest: 'I know it when I see it, but I could not define it precisely if my regulatory examination depended on it'.

This definitional ambiguity is not a trivial problem. For a risk professional, the definition matters enormously, because it determines what you are regulating, what controls are appropriate, who is accountable when something goes wrong, and which regulatory obligations apply. Risk professionals focus on and are used to working with uncertainty; however, they need both methodological and definitional certainty to be effective in their work.

An organization that treats all AI as a single undifferentiated category will end up with governance that fits none of it particularly well. Enterprises that conflate a simple rule-based decision engine with a large language model capable of generating novel text and taking autonomous actions will fail to recognize that the governance requirements for these two systems are not merely different in degree but entirely different in kind.

So before we can talk about how to manage AI risk, we need to talk about what AI is and gain greater clarity in how it operates. Not in the sense of settling a technical debate (there is genuine academic disagreement and there always has been) aimed at demarcating the boundaries of the term. But in the sense of establishing a working taxonomy that gives risk professionals the language to think clearly, ask the right questions, and build governance structures that match the actual characteristics of the systems they are trying to govern.

The Problem with the Term 'AI'

Artificial intelligence has been used as a term since the 1950s, when the computer scientist John McCarthy coined it to describe the aspiration of machines that could simulate human intelligence. For most of the intervening decades, it remained primarily an academic concept: occasionally exciting, frequently over-promised, and persistently disappointing in practice. The pattern was so well established that researchers coined the term 'AI winter' (Hendler 2008) to describe the periodic collapses in funding and interest that followed each wave of inflated expectation.

What changed was the convergence of vastly increased computational power, dramatically larger datasets, and a set of algorithmic breakthroughs that made it possible to build systems capable of recognizing complex patterns in language, images, and other data at a level approaching or exceeding human performance. The development of deep learning architectures in the 2010s, and the emergence of transformer-based large language models (LLMs) in the late 2010s and early 2020s, transformed AI from a research aspiration into a deployed technology affecting millions of people in their daily lives.

The word 'AI' now covers an enormous range of things: the spam filter in your email client, the recommendation algorithm on a streaming platform, the credit scoring model used by a bank, the chatbot on a customer service portal, the large language model producing this paragraph's equivalent in 10,000 organizations simultaneously, and the experimental agentic systems capable of writing code and executing multi-step tasks with minimal human direction. These systems share a family resemblance, in that they all involve machines performing tasks that previously required human-like intelligence; but they operate on completely different principles requiring vastly different governance responses.

The risk professional who lumps them together is not doing risk management. They are doing something that looks like risk management from a distance but will fail the moment a real problem arises.

A Working Taxonomy for Risk Professionals

For the purposes of EW-AiRM™ and this book, I use a taxonomy that distinguishes five categories of AI systems and three deployment categories covering some of their usage, increasingly device-embedded systems, and the emergence of quantum computing. These categories are not exhaustive, and the boundaries between them are not always crisp, because many deployed systems combine elements of more than one. But they capture the distinctions that matter for governance (Tables 1.1–1.5).

Neurosymbolic AI (see Table 1.5) is positioned as the fifth system category to signal an architectural pathway forward. Unlike the prior four categories, which pose escalating governance challenges (opacity, non-determinism, action-before-review, systemic concentration), neurosymbolic AI demonstrates how hybrid neural-symbolic design can mitigate those challenges through explicit reasoning, verifiable constraints, and human-interpretable decision logic.

Table 1.1 Traditional ML and deep learning definition

Traditional ML and Deep Learning	Supervised learning systems trained on labelled historical data to learn patterns and produce classifications, predictions, or probability scores. Includes fraud detection, credit scoring, insurance underwriting, and document classification (traditional ML), as well as image and speech recognition (deep learning). Outputs are reproducible given fixed model parameters. Traditional ML governance frameworks are mature; deep learning governance frameworks are evolving. Interpretability and explainability challenges increase with model complexity.

Table 1.2 Generative AI definition

Generative AI	Probabilistic systems that generate statistically plausible content, such as text, images, audio and code, by learning patterns from vast training datasets. Includes large language models such as GPT-5, Gemini, Claude, and Llama. Outputs are non-deterministic and may include hallucinations (factually incorrect or fabricated statements appearing plausible). Behaviour evolves with model updates, retraining, and prompt engineering. The non-deterministic nature and opacity of reasoning chains demand governance approaches centred on transparency, human oversight, and continuous monitoring, which are different from traditional ML risk frameworks.

Table 1.3 Agentic AI definition

Agentic AI	Systems designed to plan and execute sequences of actions toward specified objectives, often with delayed or post-hoc human oversight rather than pre-action approval. Capabilities include accessing external systems (APIs, databases, web services), executing code, modifying data or configurations, and interacting with other automated systems. The governance challenge is qualitatively distinct: the system acts in the production environment before humans can intervene. This inverts the control model from approval-before-action to monitoring-and-correction, placing heightened demands on HAiPECR™ considerations such as human oversight (H), accountability (A), and conduct risk management (C). Examples include OpenAI's agent mode (formerly Operator), Claude in agentic mode, and broader AI agent frameworks, with the potential for unintended cascading effects, especially in multi-agent or interconnected environments, which raises AI Black Swan governance considerations.

While currently research-dominant, its alignment with UNESCO 2025 Neurotech Recommendation principles on transparency and human oversight positions it as a governance-enabling architecture for regulated organizations preparing high-assurance AI deployment. The EW-AiRM framework addresses its governance requirements in Pillar 3 (Data and Technological Maturity), Pillar 5 (Governance and Accountability), and the Operational Layer (knowledge governance, rule validation, hybrid failure modes).

Table 1.4 Foundation models definition

Foundation Models	Large pre-trained models, such as GPT-5, Gemini, Claude, Llama, and Mistral, that encode learned patterns across diverse domains and serve as reusable infrastructure for downstream applications. Most organizations access foundation models via APIs, cloud platform integrations (Microsoft Copilot, Google Vertex AI, Virtual Machines), or licensed on-premise instances, rather than training models from scratch. This creates systemic concentration risk: if a foundation model becomes unavailable (technical outage, vendor policy change, API deprecation), exhibits unexpected behaviour shifts (model versioning, fine-tuning drift, adversarial inputs), or is withdrawn (licensing change, geopolitical restriction), all applications built on it are affected simultaneously. The opacity of foundation model reasoning, limited visibility into behaviour changes, and asymmetric dependence on vendor decisions create governance challenges across HAiPECR's human oversight (H), accountability (A), and resilience (R) dimensions.

Table 1.5 Neurosymbolic AI definition

Neurosymbolic AI	Systems that combine neural network architectures with symbolic reasoning represented by means of rules, knowledge graphs, logical constraints, producing AI that is simultaneously more capable and more interpretable than pure neural approaches. Examples include IBM's and SRI's neurosymbolic research programmes, and emerging platforms. By constraining neural outputs with logical rules and knowledge structures, neurosymbolic AI addresses the explainability and accountability challenges that plague pure neural and generative AI in regulated domains. Current deployments are limited, primarily in research and early enterprise pilots. However, the architecture aligns with UNESCO 2025 Neurotech Recommendation principles on transparency and human oversight, making it governance-relevant for organizations anticipating AI deployment in high-assurance, regulated use cases.

The following entries are not standalone new AI system types. Instead, they are defined by how AI is used (Shadow AI, Table 1.6) or by the contextual setting in which AI is deployed as an activity enabler (Quantum Computing). Both also represent EW-AiRM Governance Threats.

Before introducing Hybrid quantum-classical AI systems (see Table 1.7), the more immediate governance concern is the threat quantum computing poses to cryptographic security. Sufficiently powerful quantum computers will break current asymmetric encryption (RSA, ECC), compromising the confidentiality of encrypted data and the authenticity of digital signatures.

This creates systemic risk for all AI systems that depend on cryptographic security for API communication, data protection, and supply chain integrity. This threat is detailed in Pillar 3

Table 1.6 Shadow AI definition

Shadow AI	The use of external, consumer-grade AI tools (e.g. ChatGPT, Gemini, Claude subscriptions) by employees to process organizational or sensitive data, outside the organization's approved systems, data governance frameworks, and vendor agreements. Shadow AI typically exposes data to external processing, unknown data retention and use policies, lack of contractual protections (data processing agreements, residency requirements), and vendor controls that fall outside regulatory compliance frameworks (GDPR, SOX, industry-specific rules). The result is opaque, unauditable use creating material risk across data governance, regulatory compliance, intellectual property protection, and employee conduct. Shadow AI occurs at scale in most organizations, often undetected. Governance response requires visibility mechanisms (tool discovery, usage monitoring), policy clarity, and user enablement with approved alternatives.

Table 1.7 Hybrid quantum-classical AI systems definition

Hybrid quantum-classical AI Systems	Hybrid quantum-classical AI represents an emerging deployment category by combining classical ML with quantum computing for molecular simulation, portfolio optimization, and specific optimization problems. These are usually accessible through cloud platforms (AWS Braket, Azure Quantum, Google Quantum AI, IBM Quantum). Current deployments remain limited to research and early-stage enterprise pilots in pharma, materials science, and financial services. However, governance frameworks that ignore quantum-classical AI will face maturity gaps. More critically, quantum computing poses a distinct and near-term cryptographic threat to classical AI infrastructure: sufficiently powerful quantum computers will break current asymmetric encryption (RSA, ECC), compromising confidentiality of encrypted data and authenticity of digital signatures. This threatens all AI systems dependent on cryptographic security for API communication, data protection, and supply chain integrity. Post-quantum cryptography migration is a material governance action item for the Resilience Layer (Chapter 14) and Pillar 3 (Chapter 6, Cryptographic Readiness).

(Chapter 6, Cryptographic Readiness) and the Resilience Layer (Chapter 14, Black Swan Governance).

Each of these categories presents a distinct governance challenge and as new innovative approaches such as neurosymbolic AI (see Table 1.5) develop further, it is worthwhile including some of those as part of EW-AiRM's emerging risk horizon scanning.

The remainder of this chapter addresses the two most consequential distinctions for risk professionals: the unique governance challenges of generative AI, and the particular danger of shadow AI.

Why Generative AI Is Different

The most important thing a risk professional needs to understand about generative AI is that it is not a more capable version of traditional software. It is a qualitatively different kind of system that requires a fundamentally different approach to governance.

Traditional software is deterministic, meaning that when it is given a particular input, it produces a resulting output that is consistently reproducible, and in accordance with rules that can be inspected, tested, and documented. Further, if it produces a wrong output, you can trace the error by identifying the line of code, the data error, and its logical flaw. Such system bugs can be found and fixed, hence there's a level of outcome certainty based on the inputs.

In stark contrast to this, generative AI is probabilistic, meaning it produces outputs based on statistical patterns learned from enormous training datasets. This is the result of systems generating outputs that are statistically consistent with the patterns they have learned, but they are not reproducible to match each output. Quite the opposite, two identical prompts can produce different responses and the same system can perform brilliantly on one task and fail entirely on a closely related one. Most worrying for risk professionals is that such systems have no reliable internal mechanisms for distinguishing between what they know and what they are confabulating.

Hallucination: The Feature Nobody Wants to Talk About

Hallucination, the generation of confident, plausible-sounding outputs that are factually incorrect, is the characteristic of generative AI that most immediately concerns risk professionals. And it should concern them, because this is an intrinsic feature of how probabilistic language models work and cannot be simply engineered away.

A language model generates text by predicting, at each step, what token (roughly, what word or word fragment) is most likely to follow the ones that came before, given its training. This process produces text that is statistically similar to human writing, presenting the outputs as confident, well-structured, and authoritative. But statistical similarity alone is not the same thing as accuracy. The model does not have a separate truth-checking mechanism built-in, nor does it know when it is making things up. It simply generates plausible continuations based on statistical inference.

The governance consequence is significant. You cannot treat a generative AI output with the same trust as a traditional system output. If a traditional credit scoring model says the score is 720, that number was computed from documented inputs using a documented formula. If a large language model produces a research summary with citations, those citations may or may not exist, and you cannot rely on the summary accurately representing the sources cited. The model cannot tell you which parts of its output are reliable and which parts are invented.

I have used AI myself, extensively, for a long time. In preparing materials for courses and for this book, I have used it to research, to draft, to test ideas. And I have learned, through repeated experience, that you must verify everything. The model produces outputs that **look** correct (i.e. they display in the right format, vocabulary, and structure) far more reliably than it produces outputs that **are** correct. This is an important distinction worth keeping in mind.

One of the practitioners in the training sessions that informed this book described the experience of using an AI research tool under time pressure to draft a client-facing document. The document contained factual errors: incorrect figures, a source that did not exist. The document went to the client. The consequences were serious. 'It looked right', she said. 'That was the problem. It looked exactly right'. And, as it turned out, it was too good to be true!

Model Collapse and Degradation

A related risk, less immediately visible but potentially more systemic, is model collapse. As AI-generated content proliferates across the internet (e.g. in news articles, social media and research papers), and countless other sources, it is increasingly becoming part of the training data for subsequent AI models. Models trained on AI-generated content tend to produce outputs of progressively lower quality, as the statistical patterns they learn become increasingly circular.

This is a documented and well-researched phenomenon and not just a theoretical risk. The implications for organizations that depend on AI-generated research, analysis, or content creation are significant and are not yet fully understood, but with likely implications both from a reputational risk and legal risk perspective, let alone the risk of losing customers' trust.

Probabilistic Systems and Governance

The probabilistic nature of generative AI has a further governance consequence that is less often discussed: it means that the system you validated yesterday is not necessarily the system you are running today. Foundation model providers update their models, including minor updates, and sometimes significant changes, on a continuous basis. The model that passed your validation suite last quarter may be producing considerably different outputs now, without any action on your part, no explicit notification or model versioning of your own systems.

This is a structural incompatibility with traditional model risk governance, which assumes that a validated model remains in that approved state until someone explicitly updates it. In the generative AI world, that assumption no longer holds. Through-the-lifecycle monitoring (Pillar Six of the EW-AiRM framework) exists precisely to address this problem.

Shadow AI: The Risk You Cannot See

Among near-term AI governance challenges, shadow AI presents an immediate, and at the time of writing, entirely impossible to monitor and hence, invisible risk. Shadow AI occurs when employees use personal or consumer AI tools to process unauthorized and confidential organizational data. They may use such tools with good intentions, for instance by uploading a client document to ChatGPT for summarization, pasting confidential strategy reports into Gemini for key-point extraction, or using a personal Claude subscription to help draft regulatory submissions.

In each case, organizational data (that will be confidential, commercially sensitive, and subject to legal data protection and retention obligations) is transmitted to external AI services operating in jurisdictions beyond the organization's control.

Shadow AI breaches the Strategic Layer's Pillar 5 (Governance and Accountability), exposes data in violation of its Pillar 3 (Data and Technological Maturity), and represents uncontrolled conduct risk (C dimension of HAiPECR). Governance response requires visibility mechanisms (tool discovery, usage monitoring), policy clarity, embedding the right organizational culture and user enablement with approved alternatives.

Notably, from my own experiences, shadow AI appears ubiquitous, not marginal and, in every one of my training sessions, participants reported it occurring in their organizations and often at significant scale.

The root is not malice, often quite the opposite, but is closely related to conduct risk and organizationally misaligned incentives: employees tend to adopt personal AI tools because they outperform approved organizational alternatives, and workplace pressures reward speed over compliance. The governance challenge is therefore twofold: close staff's capability gap and address the implicit incentives that make shadow AI the rational choice.

Shadow AI reflects **three interconnected AI governance gaps**:

- **Low AI literacy:** employees do not fully understand that uploading organizational data to consumer services constitutes regulated data transfer with significant compliance implications.
- **Capability mismatch:** when organizational tools are inferior to personal alternatives employees may rationally choose the superior option.
- **Policy vacuum:** AI use governance frameworks, where they exist, are often vague, poorly communicated, and unmonitored.

The data leakage risk from shadow AI is significant and qualitatively different from traditional data exfiltration scenarios. Before AI, employees who wanted to exfiltrate large volumes of confidential information had to take deliberate steps, often involving real-world (IRL) interactions including printing, downloading to USB drives, or email forwarding; all of which triggered data loss prevention (DLP) controls.

With shadow AI, an employee can upload confidential client files or entire project portfolios to consumer AI services in seconds, from any device, completely bypassing traditional data governance controls. Established organizational security perimeters have not merely been weakened; they have effectively ceased to exist. In many enterprises, shadow AI represents the first generation of data loss that is both massive in scale and invisible in execution.

Device-embedded AI

Shadow AI is fundamentally a *software-enabled* governance problem, an unauthorized employee choice to use external tools. But there is a related and more structural threat emerging from the proliferation of AI-enabled hardware devices: *device-embedded AI* (Table 1.8). Whereas shadow AI exploits the capability gap between organizational and personal software tools, device-embedded AI exploits the physical perimeter intended to govern devices that connect to work infrastructure and grant embedded AI systems direct access to sensitive data and network resources. The governance response is therefore qualitatively different and not about software policy or user enablement, but about hardware access control, network isolation, and the foundational assumption that internal systems remain separated from external devices.

Table 1.8 Device-embedded AI

Device-embedded AI	Systems that embed autonomous AI capabilities directly on local devices (handheld assistants, USB-connected hardware, edge-deployed models, and smartphones with embedded Generative AI systems) capable of interacting with external systems and networked machines. Examples include the Rabbit R1 and emerging AI-enabled handheld platforms. These devices can connect to work infrastructure via USB cable or wireless connection and, with screen-sharing or functional access permissions, grant the embedded AI system visibility and executable access to sensitive data on connected machines, which in turn store files, credentials, network connections, browser sessions, email. The governance challenge is distinct from shadow AI: device-embedded AI operates outside organizational infrastructure and control and can establish data exfiltration or lateral movement pathways that bypass traditional perimeter security controls. This threat sits at the intersection of hardware governance, physical access control, and network architecture, demanding urgent attention to device policies, USB port restrictions, and the maintenance of fundamental network isolation that security architecture has long assumed.

Device-embedded AI is not yet a widespread deployment challenge as current hardware platforms remain niche and early-stage. However, the trajectory toward ubiquity is clear, particularly as AI-enabled devices proliferate in consumer markets and thereby naturally will find their way into workplace environments. The governance gap is urgent: most organizations have adequate USB port policies aimed at preventing data transfer via USB drive and other ports, but few have governance frameworks that contemplate AI-enabled hardware with autonomous action capabilities and network access. The difference is critical, given that a USB drive is inert whereas a device-embedded AI system is active, capable of exfiltrating data at scale and traversing network boundaries to access systems and credentials beyond the initially connected machine.

The regulatory framing for device-embedded AI is emerging through the UNECE Common Regulatory Arrangement (CRA) framework and the UNECE WP.6 AI Declaration for Products and Services with Embedded AI Systems (2024). These instruments establish mutual recognition protocols for AI systems deployed in products and services that cross borders and jurisdictions. Under the CRA framework, organizations procuring or deploying devices with embedded AI are expected to assess the risks: data access, lateral movement, and cross-border data flows. The Declaration translates these principles into voluntary commitments by signing governments or government agencies: producers disclose their mitigation controls, and deployers maintain visibility into the AI capabilities and data pathways of every connected device.

Procuring an AI-enabled device is no longer a hardware choice, it is a governance decision, with documented risk assessment, data-impact analysis, and control validation behind it, informed by UNECE WP.6 CRA protocols and aligned with your organization's risk appetite and data protection obligations.

The Regulatory Landscape: What Risk Professionals Need to Know

AI is now subject to a growing and fragmented body of regulation. Understanding the landscape is essential, not to ensure compliance with every jurisdiction simultaneously (that is much more of a specialist legal question), but to understand the direction of travel and to ensure that the governance framework you are building will be compatible with the regulatory environment you will be operating in.

The EU AI Act

The EU AI Act is the most comprehensive AI regulation currently in force and is likely to serve as the global benchmark for AI regulation in the way that GDPR became one for data protection. Its central innovation is a risk-tiered approach that categorizes AI applications by the level of risk they pose and then applies proportionate obligations accordingly.

At the top of the hierarchy are unacceptable-risk applications, which are prohibited entirely: AI systems used for social scoring by governments, real-time biometric surveillance in public spaces, and AI systems that exploit cognitive vulnerabilities.

Below that are high-risk applications, such as AI used in critical infrastructure, education, employment, essential services, law enforcement, migration, and the administration of justice and which are subject to significant conformity assessment, transparency, human oversight, and documentation requirements.

UK and US Approaches

The UK has taken a principles-based approach rather than enacting a single AI Act. The Financial Conduct Authority (FCA), the Information Commissioner's Office (ICO), the Competition and Markets Authority (CMA), and other sector-specific regulators are developing their own AI guidance. The AI Playbook for UK Government sets expectations for public sector deployments. The Orange Book, which represents HM Treasury's risk management framework (HM Treasury 2023), is increasingly referenced in discussions of AI governance in the public sector, and EW-AiRM aligns naturally with its principles.

In the United States, the National Institute of Standards and Technology AI Risk Management Framework (NIST AI RMF) is voluntary but increasingly influential and has been referenced in regulatory and legal frameworks. It organizes AI risk management around four core functions: Govern, Map, Measure, and Manage.

The Securities and Exchange Commission (SEC) has issued guidance on AI-related disclosures for listed companies. The regulatory landscape is fragmented and, in the current political environment, likely to be uncertain for some time, but the direction of travel globally is clear: AI regulation is coming, and it will progressively tighten.

ISO 42001 and International Standards

ISO 42001, which represents the international standard for AI management systems, is the AI equivalent of ISO 27001 for information security (International Organization for Standardization (ISO) 2022). It provides a certification framework for organizations' AI

governance systems, offering an internationally recognized benchmark against which AI governance can be assessed and audited. Organizations building an AI governance framework should ensure it is compatible with ISO 42001 requirements, both to facilitate eventual certification and because the standard reflects current best thinking on what robust AI governance looks like.

The UNECE Declaration

In 2024, the United Nations Economic Commission for Europe (UNECE) Working Party on Regulatory Cooperation and Standardization Policies (WP.6) published 'Compliance of products with embedded artificial intelligence: Overarching principles and declaration to promote convergence of product regulations' (ECE/TRADE/486).

This instrument represents the leading edge of international regulatory thinking on the governance of AI embedded in physical and digital products. It comprises two distinct components: an Overarching Common Regulatory Arrangement (CRA), providing a voluntary framework for regulatory cooperation and conformity assessment of products with embedded AI; and a Declaration for technical regulation of products with embedded artificial intelligence, providing a signing mechanism through which government agencies commit to adopting the CRA's principles.

The CRA establishes a structured approach across five regulatory dimensions: setting legitimate regulatory objectives, identifying and assessing risks, identifying relevant international standards, establishing mutually recognizable conformity assessment procedures, and establishing market surveillance mechanisms.

It articulates three risk tiers (high, limited/medium, and low) with corresponding conformity assessment requirements ranging from supplier's declaration of conformity for low-risk products to independent third-party assessment for high-risk applications. The CRA references foundational international standards including ISO/IEC 42001 (AI management systems), ISO/IEC 23894:2023 (AI risk management), ISO/IEC TR 24028 (trustworthiness), ISO/IEC TR 5469:2024 (functional safety), IEC 62443 (industrial automation security), and IEEE 7001-2021 (transparency of autonomous systems).

Among the CRA's most significant contributions is the explicit articulation, at international level, of a foundational governance principle: 'In situations where the risk of an error associated with an AI system is high, human decision making shall be included wherever possible. AI systems should not be able to override human control' (UNECE 2024, p. 3). This principle is foundational to EW-AiRM and to the HAiPECR ethical filter: it operationalizes the Human Oversight (H) dimension of HAiPECR and anchors the Strategic Layer's Pillar 5 (Governance and Accountability) at the international regulatory level.

Unlike the EU AI Act, the UNECE Declaration is not directly binding. Its influence operates through voluntary adoption by government agencies, which covers initially those participating in the WTO Technical Barriers to Trade (TBT) Committee, and progressively any product-regulation authority that signs on through its role as a normative anchor for downstream standards work at ISO/IEC JTC 1/SC 42 (AI). For organizations deploying products with embedded AI across multiple jurisdictions, the CRA framework provides a procedural pathway toward regulatory consistency through outcome equivalence, without requiring uniform global rules.

This represents a practical alternative to both fragmented national regulation and impossible top-down harmonization.

The author led the development of this publication as the project lead within UNECE WP.6, working with contributors across UNECE member states, standardization bodies, and conformity assessment communities.

The Concentration of Power: A Systemic Risk

There is one dimension of the current AI landscape that deserves specific attention as a matter of systemic risk because it is both significant in scale and widely underappreciated in enterprise risk management discourse: the extraordinary concentration of AI capability in a very small number of organizations further exacerbated by substantial geopolitical concentration.

The foundation models that underpin most enterprise AI deployments today are produced by fewer than 10 organizations globally, the majority of which are headquartered in the United States: OpenAI, Google DeepMind, Anthropic, Meta, and Microsoft (via its strategic partnership with OpenAI), with smaller frontier laboratories such as xAI representing notable secondary players.

In China, a comparable parallel ecosystem has emerged, led by DeepSeek, Alibaba (Qwen), Baidu (Ernie), and ByteDance, among many others. Beyond these two ecosystems, in Europe, the United Kingdom, and across the rest of the world, the relationship with frontier AI is primarily one of consumption rather than production. Mistral AI in France remains the most prominent non-US, non-Chinese frontier laboratory; the structural reality, however, is an asymmetric dependency on American and Chinese foundation models.

This concentration creates systemic risk at three levels: capability concentration (a small number of organizations control the underlying technology powering global AI deployment); geographic concentration (frontier development is bipolar, US and China); and strategic dependency (the rest of the world is structurally dependent on technology produced under jurisdictions and policies it does not control). Each level has implications for sovereign AI policy, supply chain resilience, and the governance of cross-border AI deployment.

For risk professionals, this concentration has direct governance implications. If a foundation model experiences failure, regardless of whether through critical security vulnerability, significant model degradation, regulatory intervention, or vendor business disruption, then every enterprise dependent on that model is affected simultaneously, creating correlated failure across multiple sectors and jurisdictions.

This is a systemic risk structurally analogous to those that financial regulators address through critical financial market infrastructure (FMI) frameworks (CPSS–IOSCO 2012): single points of failure that can transmit shock across entire systems. It is addressed specifically in the Resilience Layer (Chapter 14).

The competitive dynamics in this space are extreme. Frontier laboratories operate under intense financial and competitive pressure to deploy increasingly capable systems faster than peers. Safety research consistently lags behind deployment velocity, creating a structural asymmetry between capability advancement and risk mitigation.

Risk professionals who assume that commercially deployed AI systems have undergone thorough testing and regulatory approval comparable to pharmaceutical products or civil aviation systems are operating on an assumption the evidence does not support. And governance frameworks must address such structural asymmetry explicitly, rather than presume it away.

HAiPECR: Introducing the Ethical Filter

Before moving to Chapter 2, it is important to introduce a concept that will appear throughout this book and that is central to the EW-AiRM framework: HAiPECR, short for the Human-Ai Paradigm for Ethics, Conduct, and Risk.

Conveniently, the acronym also maps to seven thematic dimensions deliberately aligned to the 10 core principles for the Ethics of AI set out in the UNESCO Recommendation on the Ethics of Artificial Intelligence (UNESCO 2021), and grounded in that Recommendation's four foundational values (Human rights and human dignity; Living in peaceful, just, and interconnected societies; Ensuring diversity and inclusiveness; and Environment and ecosystem flourishing).

These seven HAiPECR dimensions are:

- **H – Human oversight and autonomy,** asking if effective human control is preserved, and if a human override or kill-switch is embedded at all lifecycle stages?
- **A – Accountability:** challenging whether there is a clear, enforceable accountability chain, with appropriate oversight and audit, complemented by whistleblower protections?
- **i – inclusivity and participation,** querying if the system has been designed inclusively, with diverse multi-stakeholder consultation across the lifecycle, which is a prerequisite that infuses every other dimension rather than standing apart from them.
- **P – Protection of data, privacy, and neural integrity,** probing if personal data, privacy, and for neurotechnology, the human brain and freedom of thought are respected and protected?
- **E – Ethics, comprising transparency, explainability and fairness,** evaluating if the system is free from bias and discrimination, and whether its logic and decisions can be understood by those affected?
- **C – Conduct risk, covering both AI and human behaviour,** probing whether the system's use is proportionate and necessary, and whether the conduct incentives it creates for human operators and users are acceptable?
- **R – Rights and non-harm, including intergenerational rights and environmental sustainability,** asking if the system avoids individual and collective harm, and if its sustainability implications are governed as a matter of rights owed to present and future generations?

HAiPECR originated from my earlier research in conduct risk thinking, specifically in the recognition that how human beings interact with AI systems has profound implications for ethical behaviour, for regulatory compliance, and for the cultural norms of organizations. It was developed at the intersection of the UN work on AI regulation and the earlier conduct risk paradigm work, and it is now hosted on the OECD.AI tool website (https://oecd.ai/en/catalogue/tools) and freely available to practitioners globally.

It appears throughout this book because ethics in AI governance is not a separate topic that can be addressed in a single chapter and then set aside. It is an accompanying ethics dimension of every topic: of how you assess strategic alignment, of how you define risk appetite, of how you govern AI deployments, of how you monitor systems through their lifecycle, and of how you respond when things go wrong.

HAiPECR is the mechanism for keeping that ethical dimension present and active throughout the governance process, including across all three EW-AiRM layers, rather than treating it as an afterthought or a compliance box.

Definitions That Matter

To close this chapter, a brief working reference for the terms that will appear most frequently in what follows. These definitions draw on the UNECE WP.6 CRA and Declaration, the NIST AI RMF, the EU AI Act, and the work of Professor Stuart Russell, whose formulation of the human-centric AI principle has been particularly influential in the development of EW-AiRM (Tables 1.9–1.16).

With this taxonomy in place, we can move to Chapter 2 and focus on the next question 'why are the frameworks we currently have not enough to govern it?'

Table 1.9 AI system

AI System	An engineered system that generates outputs (e.g. content, predictions, recommendations, or decisions) for a given set of human-defined objectives. (UNECE / Stuart Russell / NIST).

Table 1.10 AI risk

AI Risk	The measure of probability and impact of an AI-related threat affecting an organization's operations, assets, individuals, or broader society. Includes risks from intended use, unintended use, misuse, and emergent behaviour.

Table 1.11 AI governance

AI Governance	The strategic framework of policies, structures, and processes that ensure AI systems are aligned with organizational objectives, ethical principles, and legal requirements.

Table 1.12 AI safety

AI Safety	Practices aimed at mitigating risks from unintentional AI behaviours: software errors, misaligned objectives, emergent behaviours, and model failures. Distinct from AI Security.

Table 1.13 AI security

AI Security	Measures to protect AI systems against intentional adversarial actions: prompt injection, data poisoning, model theft, and adversarial attacks.

Table 1.14 Explainability (XAI)

Explainability (XAI)	The ability to provide a human-understandable justification for why an AI system reached a specific output. Increasingly required by regulators for high-risk AI applications.

Table 1.15 Algorithmic bias

Algorithmic Bias	Systematic errors in an AI system that create unfair outcomes for certain groups, typically arising from historical patterns in training data or from flawed objective specifications.

Table 1.16 High-risk AI

High-risk AI	AI applications that profile individuals or influence critical decisions in employment, lending, healthcare, law enforcement, or other domains where errors significantly affect people's rights or life chances. Defined under the EU AI Act and analogous regulatory frameworks.

Practitioner Questions

Before moving on, consider these questions in the context of your own organization:

- *Can you identify every AI system currently operating in your organization and categorize it using the five-type taxonomy above? Not just the ones sanctioned by IT, but all of them, including those being used by individual employees on personal devices?*
- *Does your organization have a documented definition of 'AI' for governance purposes? Does that definition distinguish between traditional ML, generative AI, agentic AI, and foundation model dependencies?*
- *Do you have any visibility into shadow AI use in your organization, i.e. the tools employees are using on personal accounts to process organizational data? If not, what would it take to get that visibility?*
- *Have you assessed which of your organization's AI applications would be classified as 'high-risk' under the EU AI Act? Do you know what compliance obligations attach to those applications?*
- *Is there a member of your board or executive committee who could, if asked by a regulator, give a confident, substantive account of the AI systems your organization deploys, the risks they create, and how those risks are governed?*