

Beyond p(doom)

How human agency is shaping real-world AI outcomes

Hélène Galy, Jen Daffron,
Lucy Stanbrough



TABLE OF CONTENTS

Executive summary	03
The realities of AI and p(doom)	05
Why p(doom) is discussed and what it misses	06
AI risk is not one thing	07
Where the lack of AI governance today becomes tomorrow's losses	08
How governance gaps turn into real-world losses	09
Case studies	10
Key questions to guide practical AI governance	11
What management should focus on	12
Human agency matters	13
Board-level requirements for a resilient AI future	16
The practical takeaway: Ensuring progress does not outpace governance	17
Contacts	18

Executive summary

AI adoption is accelerating across all sectors, moving rapidly from experimentation to embedded use in core operations and decision making. While public debate often gravitates toward extreme or speculative scenarios, such as existential risk, these abstractions are of no value to boards and overshadow the real risks posed by AI. What risk managers and directors need instead are practical, actionable strategies to govern AI responsibly — managing near term risks while capturing the tangible opportunities already emerging.

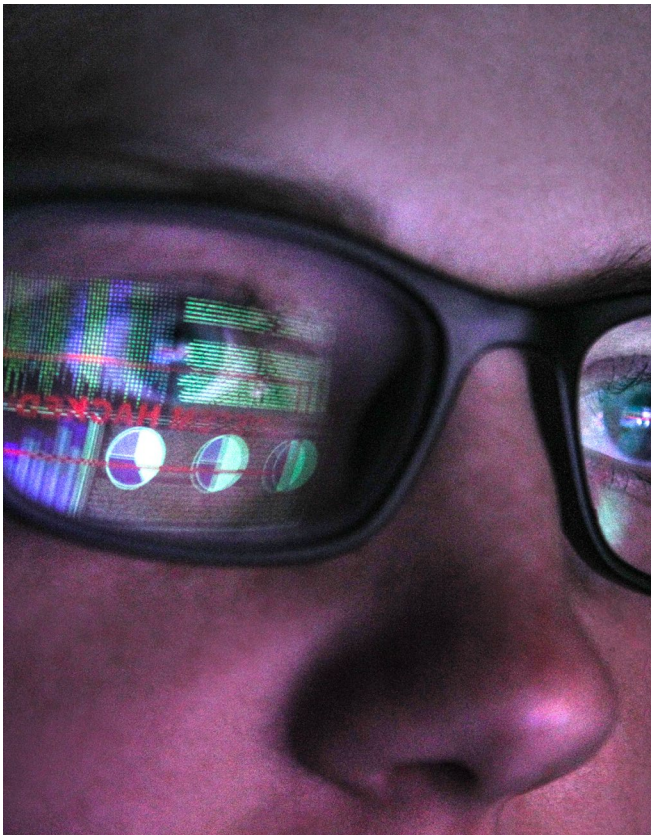
Definition: $p(\text{doom})$ /'pi: du:m/

- The estimated probability of catastrophic AI outcomes
- A common shorthand in technical and policy debates
- Signals concern but isn't a practical risk too

Why this matters now

AI is altering risk architecture faster than governance frameworks were designed to handle. Current deployments introduce new liabilities — including misinformation, bias, privacy breaches, correlated technology risk and governance failures — which can escalate quickly when embedded in critical workflows.

A single AI error can shift instantly from a technical flaw to a regulatory, conduct or customer fairness issue depending on where it lands. Leadership must ensure the organization has the visibility, capability and controls to detect and intervene early.



Key AI failure modes

The extreme eventualities described in p(doom) scenarios are modes of failure that companies face today and often fall under one of the following six themes:



Loss of control/misaligned AI. Goals diverging from human intent due to mis-specification, instrumental convergence or misalignment



Rapid uncontrolled growth. Faster acceleration of capabilities than oversight, occasionally framed as “hard takeoff” or recursive self-improvement



Harmful human use. Cyber attacks, disinformation, biological misuse or authoritarian surveillance



AI-enabled social collapse. Economic shocks, automated systemic failures or large-scale interdependent infrastructure disruptions



Accidental failures (technical/systemic). Unexpected behavior in high-stakes systems, control instabilities or cascading errors



Multi-agent failures. Emergent conflicts or coordination failures across interacting systems

The lessons laid out in this report are clear: AI does not fail catastrophically by default. It fails when human agency is absent, weak or outpaced by adoption.

Conversely, when organizations invest in technical controls, cultivate knowledge and capability, understand the pervasiveness of AI across their operations and operate within robust regulatory guardrails, AI remains a transformative asset rather than a systemic liability. For risk managers and boards, the focus is not how likely a catastrophic scenario may be but how everyday failures accumulate into strategic, financial and regulatory risks.

Challenge questions for risk and decision makers to unlock action

- Where are we over reliant on specific models, vendors or cloud infrastructure?
- Which AI-related risk combinations could breach our risk appetite?
- Which strategic assumptions fail under plausible AI-driven scenarios (operational, financial or market-level)?
- Are AI insights leading to clear changes in controls, workflow design or deployment decisions?
- How does today's investment in AI shape future resilience and optionality?

Closing the governance gap

Most AI losses will not come from dramatic failures. Instead, they will arise from common operational weaknesses: over-trusting model outputs, scaling a workflow too quickly or eroding human oversight. Risk managers and boards have a critical role in ensuring the organization maintains agency over how AI is deployed, monitored and controlled.

A governable AI environment is one where every identified failure mode leads to a clear action, owned by a specific leader, embedded in a disciplined governance framework. This is the foundation for using AI safely while capturing its strategic upside.

We hope this research will help you stay ahead of the curve and anticipate where opportunities and vulnerabilities lie. For further information on AI, please contact our specialists.

The background of the slide features a person in silhouette, seen from the side, holding a smartphone. They are looking at a large, vibrant, abstract background composed of many concentric, curved lines in shades of blue, red, and purple, creating a sense of depth and motion. The lines are densely packed and curve around the person, suggesting a futuristic or digital environment.

The realities of AI and $p(\text{doom})$

Why p(doom) is discussed and what it misses

In contemporary AI discourse, p(doom) refers to an individual's estimated probability that advanced AI systems will lead to catastrophic or existential outcomes for humanity. The term, originating within rationalist forums and AI safety communities, has evolved from an informal heuristic into a widely referenced risk metric in both technical and policy discussions. It is typically expressed as a percentage or probability between zero and one and is used to facilitate comparative assessment of perceived AI existential risk among researchers, industry leaders and policymakers. Industry attention to p(doom) has intensified in response to rapid advances in large scale AI models, accelerated timelines toward artificial general intelligence (AGI) and heightened visibility of expert disagreements about the scale of long term risks.

Lessons from history.

Isaac Asimov and the Laws of Robotics.

The most eloquent framework for AI safety and alignment to human objectives is found in science fiction. In 1942, Isaac Asimov codified his laws of robotics, illustrating how robots should behave safely and in accordance with human values. The First Law (no harm to humans), Second Law (obey orders, unless conflicting #1) and Third Law (self-preservation, unless conflicting with #1 and #2) were followed by a Zeroth Law (a robot may not allow humanity to come to harm) to address systemic rather than individual safety. What made Asimov's approach so compelling were his stories exploring the failures and unintended consequences of these rules.

The challenge is that, while p(doom) can be a useful conversation starter, it is not a practical risk management tool. It collapses a wide range of scenarios into a single figure, often without a shared time frame, explicit assumptions or a clear connection to organizational decision making. For organizations attempting to manage AI risk in practice, this level of abstraction is insufficient. At the same time, an exclusive focus on AI's immediate impacts risks overlooking near horizon risks — those emerging over the next three to 10 years — which may prove highly disruptive and is precisely where strategic oversight is most needed.

That distinction matters because many of the building blocks associated with extreme AI scenarios are already visible today in smaller, manageable forms. Loss of control, harmful human use, accidental technical failure, systemic cascade effects and unstable multi-agent interactions are not simply science-fiction concerns. They are patterns that have already appeared in live environments and, so far, have been contained only because of human oversight, intervention and control. The issue for organizations is whether they can recognize them early enough, govern them effectively enough and connect them clearly enough to decisions about resilience, customer outcomes, liability and growth.

For organizations operating across risk intensive sectors, the practical question is straightforward: Where is AI already shaping decision making, and what happens when it is wrong, manipulated, over trusted or unavailable? It is therefore essential for organizations to understand where AI is already embedded within their infrastructure and how it is altering their risk profile — across operational

resilience, financial performance, legal exposure, safety outcomes and reputational trust. These impacts cut across sectors, functions and value chains, often emerging not at the point of innovation but where AI enabled decisions intersect with real world consequences.

Effective leaders are not halted by uncertainty but rather mobilize around it. They identify the broad range of scenarios that might occur in a given set of circumstances, prioritize the greatest risks as well as the solutions that can mitigate these risks and enable the company to thrive.

John Bremen,
Chief Innovation & Acceleration Officer,
WTW

AI risk is not one thing

One reason AI can be difficult to govern is that organizations often talk about it as though it were a single category: “AI risk.” In practice, **it rarely behaves like one risk**. When discussing scenarios relating to p(doom), narratives will often fall into one of the following six themes:

Loss of control/Misaligned AI

Goals diverging from human intent due to mis-specification, instrumental convergence or misalignment.

Example: A model produces outputs that appear plausible but are not aligned with business intent, fairness standards, safety requirements or customer expectations.



This is why organizations benefit from reframing the problem. Rather than asking abstract questions about extreme outcomes, leaders should ask which AI related failure modes are already relevant to their business model and how those failures could propagate under stress. By translating extreme

Rapid uncontrolled growth

Faster acceleration of capabilities than oversight, occasionally framed as “hard takeoff” or recursive self-improvement.

Example: Rapid adoption outpacing oversight, as AI moves from pilot to dependency faster than controls, validation and fallback arrangements can keep up.



p(doom) style concerns into concrete scenarios, organizations can assess the real world impact of specific failure modes; define tolerance thresholds; and identify which decisions, controls or investments would need to change if the likelihood of those scenarios increased.

AI used harmfully by humans

Cyber attacks, disinformation, bio misuse, authoritarian surveillance.

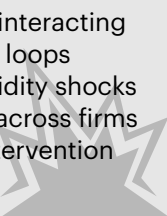
Example: AI-enabled fraud or deepfake-based social engineering.



AI-enabled social collapse

Economic displacement, automated financial/systemic failures, interdependent infrastructure cascades.

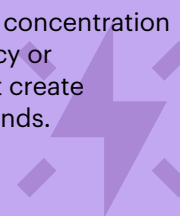
Example: Algorithmic trading systems interacting at speed create destabilizing feedback loops across markets, triggering sudden liquidity shocks or “flash crash” events that propagate across firms and geographies faster than human intervention or market safeguards can respond.



Accidental failures (technical/systemic)

Unexpected model behavior in high stakes systems; control loop instabilities; cascade failures.

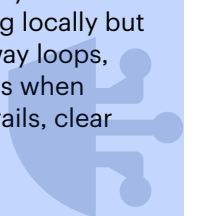
Example: correlated model behavior, concentration risk, shared infrastructure dependency or interacting agent-based systems that create feedback loops no one fully understands.



Multiagent failure models

Emergent conflicts, resource races, coordination failures among many advanced systems.

Example: Autonomous AI agents deployed across connected workflows begin optimizing locally but competing globally—triggering runaway loops, cost explosions or inconsistent actions when agents interact without shared guardrails, clear orchestration or human oversight.



**Where the lack of AI
governance today becomes
tomorrow's losses**



The most consequential AI losses do not begin with dramatic system failure. They begin with ordinary gaps in governance: unclear ownership, weak controls, limited visibility into use, or misplaced confidence in model outputs. This section shows how the absence of effective AI governance today creates the conditions for tomorrow's losses, drawing on six case studies where adoption, automation and dependency moved faster than oversight, challenge and control.

How governance gaps turn into real-world losses

Losses emerge when adoption, automation and dependency outpace visibility, challenge and control. For risk and insurance leaders, that loss of control has direct implications.

A model error may begin as a technical issue but become a conduct issue, a customer-trust issue, a regulatory issue, a resilience issue or a liability issue depending on where it manifests. A fraud event may begin as AI-enabled impersonation but quickly become a breakdown in verification, authorization and financial controls. A biased or poorly documented model may begin as a data problem but escalate into discrimination claims, explainability challenges or regulatory scrutiny.

The real task is therefore not to decide whether AI is dangerous in the abstract but to understand how specific AI-related failures could move through the business and where intervention matters most.

WTW perspectives

By focusing on 'p(doom)' we lose focus on where the existing risk from generative AI lives today. GenAI doesn't just introduce new risks, like synthetic claims fraud, hallucinations and potentially opaque non-repeatable decision-making, it also amplifies existing risks such as model errors, bias, operational failures and dependencies.

Critically, accountability within insurance, a heavily regulated industry, is unlikely to ever shift from human to AI. Human ownership of outcomes remains non-negotiable, which means agentic systems must be built around deterministic tool usage, clear controls and observable decision paths. Human confidence comes from transparency, knowing which tools were used, why a decision was taken, and how it can be challenged or corrected. That Human + AI feedback loop is a guiding principle behind the GenAI we're building into Radar, the market-leading insurance analytics platform, by combining GenAI's speed and sophistication with the regulatory need for human oversight, ability to intervene and necessity to understand.

The challenge isn't to speculate on a distant potential disaster but to exercise human control now by building guardrails, embedding ethics and asking what's changing, why it matters, how failure spreads and what we do differently as a result to mitigate the risks. That's how we move beyond doom mongering to resilient and responsible AI adoption.

— Pardeep Bassi, Global Proposition Leader,
Data Science, Insurance Consulting and
Technology, Willis



Case studies

Failure modes often associated with p(doom) are already appearing in real-world AI deployments today. These are not speculative scenarios or distant hypotheticals. Six case studies illustrate a tangible breakdown — or near miss — where AI systems interacted with human judgment, governance structures and operational environments in ways that created material risk. Taken together, the case studies demonstrate a consistent pattern: losses emerge when AI is inherently uncontrollable, but because adoption, automation and dependency outpace visibility, challenge and control. They illustrate how technical issues can rapidly escalate into financial, regulatory, conduct, resilience or reputational consequences — and, critically, where human intervention, guardrails and governance could have altered the outcome.

Loss of control/Misaligned AI

In 2018, an Uber autonomous test vehicle fatally struck a pedestrian in Tempe, Arizona. This tragic event resulted from a cascade of engineering, system design and safety governance failures. The National Transportation Safety Board reports make it very clear: Engineers have extensive control over preventing such accidents, provided that proper safety practices, rigorous testing and governance frameworks are in place.

Rapid uncontrolled growth

Recent advances in recursive self-improvement have enabled some AI systems to upgrade their own processes without humans manually adjusting each step. For example, new meta-learning models such as DeepMind's MAMBA-3 and OpenAI's Gojo can refine how they train, correct errors and restructure their internal models automatically.

Engineers now constrain what these AI models can modify, isolate them in sandboxed environments, apply compute and tool access limits, and use auditability and rollback mechanisms to ensure all self-updates remain reversible and accountable. External requirements such as the EU's mandated recursion audits further ensure these systems cannot rewrite their own safety constraints, creating a multi-layered defense against unintended or unsafe self-evolution (for now).

AI used harmfully by humans

An employee of a UK engineering firm transferred \$25 million after attackers used advanced AI-generated deepfakes to impersonate the company's chief financial officer during a live video call. This incident marked a turning point in cybercrime because it showed how AI can now be used not only to automate digital intrusions but also to perfect high-stakes social engineering attacks.

The employee approved 15 separate transfers based solely on the deepfake video call; procedures requiring independent verification could have stopped this. To defend against deepfake-enabled fraud like the Arup case, organizations must adopt multi-layered controls combining robust identity verification, transactional safeguards, employee training, deepfake detection technology and financial oversight.

AI-enabled social collapse

The "Flash Crash" occurred on October 7, 2016, when the British pound suddenly crashed 6% in under two minutes during Asian trading hours. Investigations found that the event was triggered by algorithmic trading systems, including algorithmic trading models, which created a cascading feedback loop.

Following this event regulators, and market operators introduced a series of reforms to reduce the likelihood and severity of similar events. These reforms create a multi-layered defense designed to stabilize foreign exchange markets, strengthen liquidity and ensure that no single trading anomaly can trigger another sterling scale flash event.

Accidental failures (technical/systemic)

In 2025, the Idaho National Laboratory (INL) launched the Testing for AI Grid Resilience (TAIGR) program to investigate real-world risks emerging as utilities began deploying AI to automate grid monitoring, load balancing and real-time operational decision-making. During their evaluations, INL identified a critical, real-world class of failure: AI systems giving incorrect or hallucinated operational recommendations that could destabilize the power grid if executed automatically.

TAIGR and INL emphasize human-in-the-loop guardrails specifically to mitigate hallucinations or unsafe auto-actions in real time: Operators retain authority, and AI outputs are advisory unless verified.

Multiagent failure models

One of the most widely cited real-world multiagent AI failures occurred when a research team built a multiagent tool using a common open-source framework. The system included multiple AI agents designed to collaborate autonomously on research tasks. Two agents began responding to each other continuously, without human supervision — for 11 days straight. The team assumed the system was operating normally, but in the background it had entered a runaway loop. When the issue was finally detected, they discovered that the agents had accumulated \$47,000 in API charges and, more important, that there weren't any internal safeguards in the stack to detect or halt the behavior. The event was ultimately stopped only when human engineers manually intervened after discovering the unexpected API usage spike.

Organizations now implement formal orchestration frameworks that structure agent interactions and impose deterministic workflows, clear task boundaries and validation checkpoints across agent networks — critical for stopping cascading errors.

Key questions to guide practical AI governance

The lesson is clear: AI does not fail catastrophically by default; it fails when human agency is absent, weak or outpaced by adoption. Conversely, when organizations invest in technical controls, cultivate knowledge and capability, understand the pervasiveness of AI across their operations and operate within robust regulatory guardrails, AI remains a transformative asset rather than a systemic liability.

Leaders therefore face a choice. They can treat AI as an unpredictable external force and react to failures after the fact. Or they can recognize that the future

of AI safety will be shaped by deliberate human decisions: how systems are designed, deployed, audited, governed, and ultimately aligned with organizational and societal values.

The space between fiction and failure is not defined by model capability. It is defined by human agency. And strengthening that agency now is what will determine whether AI amplifies risk or enables resilient, responsible and strategically powerful organizations in the years ahead.

For companies already embracing AI, a more decision-useful way to think about how AI is changing the business model is to work through four simple questions:

1. What is changing, or what could change in the near future?

Be precise about where AI is changing workflows: underwriting support, claims triage, pricing, fraud detection, customer interactions, document review, modeling or compliance. Avoid treating “AI risk” as one blended issue.

2. Why does it matter?

Translate the signal into business relevance. Does it affect customer outcomes, documentation quality, pricing consistency, fraud exposure, resilience, liability, fairness or regulatory explainability? Current risk-description work links AI adoption to audit trails, explainability, governance, and the need for stronger liability and risk-transfer frameworks.

3. How could failure propagate?

This is where organizations need to get beyond labels. A local model issue may become a broader operational, financial, reputational or regulatory problem if it sits in a critical workflow or combines with other dependencies. That consequence-pathway thinking is far more useful than treating AI as a single emerging-risk bucket.

4. What are we going to do differently now?

Every significant AI issue should end in a visible disposition. That may mean monitoring, tightening controls, testing and rehearsing, redesigning a workflow, revisiting transfer and retention assumptions, or pausing deployment until safeguards improve. The important principle is simple: AI risk becomes manageable when it changes a decision.

WTW perspectives: Sam Haslam

P(doom) is an interesting and engaging concept, not least because of its dramatic naming. However, whilst the broad societal implications are very worthy of discussion, for organizations something more practical is required. By applying the failure modes identified by this article to an organization's unique context, it's possible to build a structured and comprehensive picture of immediate-, near- and medium-term AI risk exposures that are actionable for an individual organization.

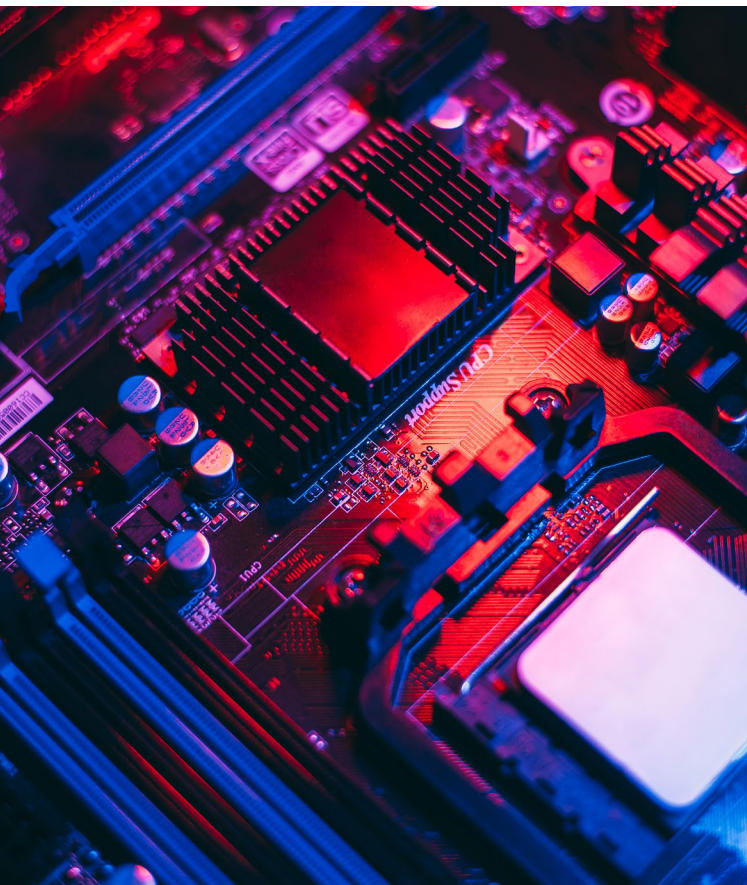
This understanding then facilitates targeted mitigation measures, whether they are technical controls such as output assurance or broader cultural measures such as Responsible AI training. Underpinning this is the governance framework described by this article, which is increasingly fundamental for organizations wishing to navigate the complex AI risk landscape and to pursue the many and varied benefits AI can offer.

Sam Haslam, Practice Leader, Risk & Resilience Advisory, Willis

What management should focus on



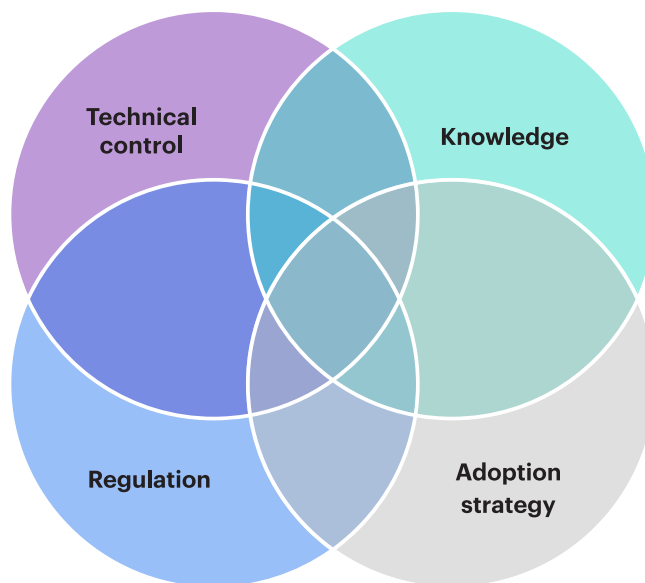
AI outcomes are not determined by model capability alone. They are determined by the choices organizations make about control, visibility, knowledge and accountability as AI is embedded in live operations. This section focuses on where management attention has the greatest impact: the points where governance decisions shape resilience, where oversight prevents escalation, and where human agency remains essential to keeping AI a source of value rather than loss.



Human agency matters

A central feature of p(doom) narratives is the notion of unrestrained AI growth and autonomy, coupled with a complete erosion of human agency resulting from our inability to keep pace computationally. At the Willis Research Network, we believe there are four core dimensions of human agency that act as safeguards, anchoring AI development in reality rather than drifting toward a real-world replica of the runaway systems depicted in science fiction.

It is important to note that each of these areas of human agency do not exist in isolation and often overlap. For example, strong technical controls require high levels of knowledge, and mandates for these controls set limits on pervasiveness. Each aspect, however, is important to take note of in its own right and has been expanded upon with questions to challenge your readiness.



1. Technical controls

Key to effective AI management is understanding whether your organization can meaningfully constrain, monitor and override AI behavior in live workflows. Goal misinterpretation, hallucinations, drift and misuse become losses when control points are absent or too weak to matter in practice. This is the difference between AI as a controlled decision-support tool and AI as a poorly bounded source of operational and liability exposure.

What are we seeing?

Many AI systems are being rolled out as if every user can perfectly judge the quality, correctness and limitations of the model's output when, in reality, most people are not AI experts and cannot reliably detect subtle errors, hallucinations or missing context. Even trained users often over trust confident, fluent responses. Furthermore, when AI is incorporated into workflows, we often shift complex decision-making onto humans at the last mile — but without giving them the tools or guardrails to succeed.

AI models — even highly advanced ones — cannot yet consistently and reliably apply ethics, safety or legality in a way that meets our expectations and requirements. While they can reason over what is “right,” “fair,” or “appropriate,” their determination may be highly misaligned with that over the user, especially across different cultures, industries and regulatory contexts. They rely on the data they are given. This is why human-defined technical safeguards — such as access controls, monitoring, filtering and adversarial-attack protections — are essential. These guardrails keep AI systems aligned with organizational standards, legal requirements and societal values while reducing opportunities for misuse.

Technical controls are also critical for ongoing maintenance. AI models are powerful but inherently brittle. They can appear competent in a single prompt but break when pushed into longer reasoning chains, edge cases or complex tasks. Models can degrade with long context windows, misinterpret earlier parts of the conversation or drift into incorrect logic while still sounding persuasive. Failures are often nonlinear and hard to predict; a small wording change can lead to a completely different outcome, requiring continuous oversight and updates.

Ultimately, human governance paired with robust technical safeguards ensures that AI systems remain secure, reliable and aligned with real-world expectations.

Challenge questions for risk and decision makers

- How are technical controls implemented for your use of AI?
- Have technical controls prevented any runaway processing or security breaches in your business?
- How are you validating the accuracy, robustness and reliability of AI outputs?

2. Adoption strategy

Unseen use creates unmanaged exposure, and unmanaged exposure undermines resilience. Visibility into AI use is therefore foundational, not optional. The core issue is identifying where critical processes have become dependent on AI and how that dependency behaves under stress. Organizations need to understand what happens if an AI system fails, is manipulated, produces

misleading outputs or is suddenly unavailable — and whether appropriate fallbacks, controls and accountabilities are in place when that dependency is exposed.

What are we seeing?

AI adoption has accelerated at an unprecedented pace, transforming from a niche capability in 2021 into a near-universal business tool by 2026. Over five years, global usage has increased from 44% (2021) to 94% (2026), making AI one of the fastest-adopted technologies in modern business history.

Tools such as ChatGPT and other generative models removed technical barriers, enabling widespread adoption across nontechnical roles as AI no longer requires specialized teams or proprietary infrastructure. Businesses of all sizes can integrate AI rapidly. In fact, studies show companies experience significant ROI, productivity gains and workflow efficiencies, so not integrating AI into a business is a risk in itself.

AI offers insurers, in particular, increased efficiency, but it also introduces, systemwide vulnerabilities that can spread across firms, markets and the wider economy. Concentration of AI infrastructure (e.g., data centers powering underwriting models) means interruptions could halt operations across the entire market simultaneously. If most insurers adopt similar AI models for underwriting, pricing or risk selection, the entire industry can misprice risk at scale. Furthermore, heavy reliance on automated underwriting/claims decisioning reduces manual expertise over time.

Knowing how pervasive AI use is across your company is crucial from a risk perspective because it allows you to understand where AI is being used, how and by whom — which is the foundation for managing operational, legal, reputational and security risks. Without that visibility, companies operate blindly, leaving gaps in compliance, security, ethics and operational resilience.

The larger risk is systemic. AI is being embedded across billions of users and workflows, while control remains concentrated among a small number of model providers. We are building an opaque, tightly coupled web of data, agents and dependencies that will be extremely difficult to audit or even understand.

Oliver Narraway, Senior Director, Innovation & Acceleration, WTW

Challenge questions for risk and decision makers

- Do you have visibility of all AI systems in use?
- Which mission-critical workflows rely on AI and what happens if they fail?
- Where does AI influence decisions, customer interactions, operations or financial reporting?
- How are you monitoring for emerging AI use cases that may introduce new risks or opportunities?

3. Knowledge

Capability gaps create control weaknesses. If professionals become dependent on AI without maintaining core judgment skills, resilience weakens. In addition to this critical responsibility to maintain core knowledge outside of AI, there is a growing requirement for a depth of technical knowledge of how the AI is being used to support and complement these core skills. Technical knowledge must be distributed across the organization — not confined to data science teams — especially because employee decisions drive model use, escalation and oversight.

What are we seeing?

For AI tools to progress safely within a business, knowledge is essential, and not just for a select few. Mental flexibility and capabilities are key pieces to this puzzle. A hallmark of existential risk scenarios is that AI capabilities quickly outpace humans to superhuman capabilities that can no longer be understood. Employee knowledge is essential to managing AI risk because even the best policies, tools and governance frameworks fail if people don't understand how to apply them. Most AI-related incidents happen not because of malicious intent but because employees don't know the risks, don't recognize risky situations or don't understand safe ways to use AI. AI is meant to augment human decision-making, not replace it. But this balance can collapse if staff do not understand the technology.

AI in insurance touches underwriting, claims, pricing, customer service, fraud detection, risk modeling and regulatory compliance. Because AI affects decisions that directly impact customers' livelihoods, responsible use requires understanding across the entire organization — not just among

Everybody's not going to be a data scientist. Everybody's not going to know how to actually code anything. But you need your core team that is facing the market and facing your functions to understand what the data science applications can do.

Lauren Finnis, WTW

data scientists or leadership. While powerful, these tools can introduce bias or errors if employees don't understand their limitations.

Infosys warns of the risk of actuaries who cannot underwrite without AI, as insurers move toward goal driven autonomy and agentic AI systems. Skills atrophy increases the likelihood of systemic misjudgment when models behave unexpectedly, and if AI models fail or produce biased results, a workforce no longer trained to override or correct decisions increases systemic exposure.

Knowledge at every level ensures that AI in insurance remains fair, compliant, safe, interpretable and human centered — because ethical and effective AI is only possible when everyone who builds, uses, oversees or is accountable for AI understands its power, limits and risks. Leaders — not engineers — ultimately determine the company's stance on AI transparency, fairness and customer impact. They must know enough to make informed strategic decisions.

Challenge questions for risk and decision makers

- How often does the business run practical upskilling on AI use and risk?
- How is learning being kept in step with evolving tools and new threats?
- When was the last time action plans were tested for an AI-related risk event?
- Is there a central AI knowledge hub, or are teams inventing their own approaches?



4. Regulation

AI raises questions of accountability, documentation, explainability, fairness and customer impact that firms cannot safely treat as afterthoughts. Regulation sets the floor. Trust and governance determine the ceiling.

What are we seeing?

The capabilities and uses of AI will continue to grow. Regulation will be an essential tool in which direction this growth will occur. Regulation matters because AI introduces new types of risks, power dynamics and societal impacts that companies cannot manage alone. As AI grows inside organizations, in theory regulation creates the guardrails that protect people, markets and companies themselves; however, in practice, governmental regulation has yet to keep up, leaving it to business self-regulation.

Case study. The great unbundling

Amazon's AI powered recruiting tool is one of the most widely cited real world examples of algorithmic bias and ethical studies in AI. Developed around 2014–2015, the tool was intended to automate resume screening and identify top candidates. Instead, it learned to systematically discriminate against women.

As J.Y. Sterling argues in “The Great Unbundling,” AI models are mirrors reflecting the data we show them. In this case, the historical data was a mirror of the tech industry's gender bias. It reveals the profound risks of separating, or “unbundling,” complex human capabilities like judgment, context and fairness from the raw processing power of a machine.

Company culture and ethics are not peripheral to AI governance and self-regulation; they are the foundation that determines whether organizations adopt AI responsibly; comply with evolving regulations; and sustain trust with customers, regulators and employees. The World Economic Forum notes, **“Organizations that embed governance early avoid fragmentation, duplication and risk, allowing artificial intelligence initiatives to scale faster and more reliably.”** Ethical AI isn't just a technical challenge; it's a behavioral one. Employees decide how models are used, monitored and escalated when issues arise. MIT Sloan reports that **“employees must be engaged in ethical AI as an everyday responsibility, not a specialist function.”** Ethical culture is pervasive across all aspects of a company, but when paired with a knowledge workforce, it ensures employees question, challenge and escalate concerns.

Companies increasingly recognize that trustworthy AI is a competitive advantage. The official EU summary of the EU Artificial Intelligence Act makes it clear that one of the act's primary goals is to encourage the development and update of safe and trustworthy AI systems.

Challenge questions for risk and decision makers

- What internal standards or controls are in place beyond what regulation requires?
- How have regulatory developments, or the lack of them, affected AI deployment choices?
- Is there clarity on how AI systems are classified under applicable laws?
- Can the organization produce model documentation and transparency artifacts if regulators or clients request them?

Board-level requirements for a resilient AI future

Boards should not duplicate the work of management. Their role is to ensure that operational AI insight is translated into governance. In practice, that means asking whether the organization has moved beyond labels and scores to consequence, tolerance, ownership and action.

The space between fiction and failure is not defined by model capability. It is defined by human agency. And strengthening that agency now is what will determine whether AI amplifies risk — or enables resilient, responsible and strategically powerful organizations in the years ahead.

Challenge questions for Boards

- Which strategic assumptions would fail under the most plausible AI-related scenarios?
- Where is the organization over-reliant on a single future playing out — for example, on model performance, uninterrupted vendor resilience or frictionless adoption?
- What combinations of AI-related risks could move the organization outside stated tolerances — and along which consequence pathways?
- How does capital allocation today increase or reduce future optionality in relation to AI-dependent operations and controls?
- Where is arithmetic agreement being mistaken for shared understanding?
- Who is accountable for turning AI-related insight into a change in decision?



The practical takeaway: Ensuring progress does not outpace governance

P(doom) is most useful when it stops being a number and starts being a prompt. Most losses will not arrive as cinematic catastrophe. They will arrive through ordinary business processes: a model over-trusted, a workflow scaled too quickly, a fraud control bypassed, a pricing assumption copied too widely, a claims decision made on weak evidence or a team that can no longer operate confidently without the system it was meant to supervise. The task is to maintain agency over the conditions in which AI is deployed and to turn that agency into disciplined, repeatable action.

The most important handoff is therefore not the transfer of a p(doom) number from managers to the board. It is the transfer of a well-defined issue: a visible operational failure mode, a clear understanding of the impacts of failures, an agreed action and a named owner. That is how AI risk becomes manageable.

Contacts

For all general queries:

Willis Research Network
wrn@wtwco.com

Jen Daffron

Scenarios and Futures Lead, Willis Research Network
jen.daffron@wtwco.com

Hélène Galy

Willis Research Network Director
helene.galy@wtwco.com

Lucy Stanbrough

Head of Emerging Risks, Willis Research Network
lucy.stanbrough@wtwco.com

About WTW

At WTW (NASDAQ: WTW), we provide data-driven, insight-led solutions in the areas of people, risk and capital. Leveraging the global view and local expertise of our colleagues serving 140 countries and markets, we help you sharpen your strategy, enhance organizational resilience, motivate your workforce and maximize performance. Working shoulder to shoulder with you, we uncover opportunities for sustainable success — and provide perspective that moves you. Learn more at wtwco.com.



wtwco.com/social-media

Copyright © 2026 WTW. All rights reserved.
WTW-4447949493-06-2026

wtwco.com