



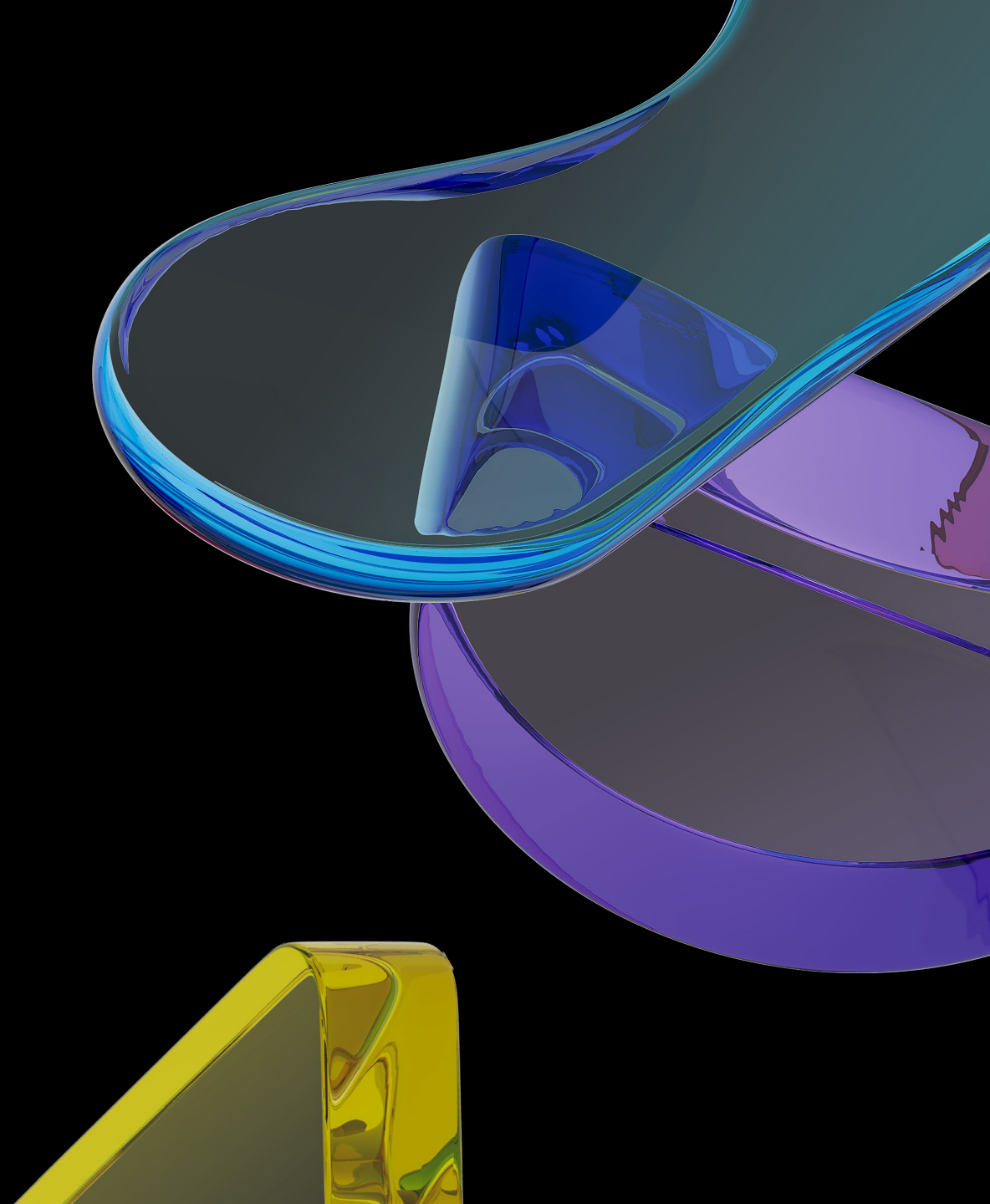
alma economics



The cost of AI failure

An economic study

October 2026



Contents

01

Foreword and
executive summary

02

Introduction and
key findings

03

State of AI
governance: leader
perspectives

04

Economic model:
methodology

05

Economic model:
findings

06

Conclusions and key
takeaways

07

Appendix

08

Your pathway to
trusted AI

Foreword

Susan Taylor Martin, Chief Executive, BSI

It is no exaggeration to say that every business leader now has AI on their agenda. Regardless of sector or size of an organization, leadership teams around the world are grappling with how to embed AI in their operations: what tools to use, how AI can enhance their offer or how to train staff to use it effectively. But as AI reshapes our businesses, economies and society more broadly, leaders also need to consider what happens when AI doesn't work as expected.

As this report shows, AI failure is not a hypothetical, but a regular, occurrence, and one that is likely to bring with it significant costs, both to individual organizations and the wider markets in which they operate. Many, if not most, organizations will experience some form of AI failure, whether that is a hallucination, the wrong information being shared with a client or a chatbot responding in an inappropriate way.

If failures are inevitable, we also need to normalize a conversation about those incidents, rather than being fearful about what happens when they are exposed. We need to encourage leaders to do everything they can to prevent failures, while being comfortable sharing what has happened when one does occur, what they have learnt, and, crucially, what additional safeguards they are putting in place to prevent a repeat occurrence. It is not one or the other – we need both preventative governance measures as well as robust review of what happened and the corrective action needed, should failure occur.

The comparison can be drawn with cybersecurity, where incidents occur with regularity, but we have gone a long way to develop a culture of disclosure when they do, with this legally mandated in some jurisdictions. Equally, taking steps to prevent issues arising is considered routine.



Just as with cybersecurity, business leaders can take action when it comes to AI. Today, organizations around the world certify to ISO/IEC 27001, demonstrating that they have an independently verified, structured approach for managing information security risks.

That same approach should be taken when it comes to AI, with leaders recognizing the imperative to install guardrails and prioritize robust AI governance. ISO/IEC 42001 provides a comprehensive framework that emphasizes ethical and responsible AI use. There are also a range of other assurance offerings to evaluate if AI models are acting as intended or adhere to emerging regulatory requirements through conformity assessments.

Failure is a natural part of life. From infancy, we make mistakes and we learn from them. The same is true of innovation; the most creative ideas are often formed from a thwarted initial effort.

We cannot expect the profound technological change we are experiencing to unfold without AI failure being a regular feature. But what happens next is a test of leadership and responsible business.

People, whether ordinary citizens or customers of a specific company, have a right to expect transparency when mistakes occur and to be confident that issues will be rectified. As AI's impact grows, good governance must sit alongside it – not to stymie innovation, but to enable this to happen in a safe way that protects us all.

There are huge benefits to be reaped as a result of AI and I do truly believe it can be a force for good. But businesses have a responsibility to their staff, their customers and society to innovate with care and consideration. Failure will happen – but leaders can determine what follows.

1. Executive Summary

Purpose and scope

BSI commissioned Alma Economics to explore the economic costs associated with AI failure, as part of its wider work on AI governance. This report details the economic study and is supplemented by polling of global business leaders around AI governance practices, including failures.

By AI failure, this research considers cases where an AI system makes an unexpected error, malfunctions, performs poorly or produces a biased output, and considers the financial consequences these failures can create for organizations and wider markets.

This study uses real-world examples of AI failure to estimate the potential scale of financial risk associated with different types of failure. It does not aim to calculate a single overall cost of AI failure, given the inherent uncertainty associated with the research.

Instead, it gives an indication of the order of magnitude of potential financial risk and shows how this can vary depending on the way AI is used and the context in which a failure occurs.





Approach

We used a case study approach in which the analysis is based on real world examples.

A broad desk-based search was undertaken to identify real-world incidents of AI failure reported in the public domain. Cases were then assessed against selection criteria, including whether they fell within the study's definition of AI failure, whether a clear monetary or otherwise quantifiable impact could be identified, and whether the failure occurred in real-world use rather than during development or testing. This resulted in seven case studies across four broad categories:

- Biased or discriminatory AI-powered decision-making;
- Hallucinations and fabricated outputs;
- Customer-facing chatbot failure;
- Agentic AI failure.

For each case study, the observed financial impact is used as a starting point for estimating the potential financial risk at a wider market level, combining four elements:

1. The severity of the observed failure;
2. The size of the relevant market;
3. The extent to which AI is used for the relevant activity;
4. Where applicable, the likelihood of that type of failure occurring.

Estimates are then scaled across geographies to provide an indication of potential global financial risk.

Headline results

Across the case studies examined, the modelling suggests that AI failure could cost approximately 0.1% to 3.6% of market revenue* each year, depending on the type of failure and the sectors or activities considered.

These estimates are anchored in observed financial consequences from real-world AI failures, including settlements, compensation payments, sanctions, refunds and excess expenditure and do not incorporate additional potential penalties under other regulatory regimes. As adoption expands and AI is used across a wider range of business activities, the scale of potential financial exposure could also increase.

The variation across these estimates reflects differences in the underlying case studies and sectors, rather than the failure categories alone. Higher estimates tend to arise where the observed failure is relatively severe, the relevant AI use is more widespread, and the estimated likelihood of failure is higher.

Conversely, even a severe individual failure can translate into a relatively low market-wide estimate where only a small share of activity is exposed or where the failure is expected to occur infrequently.

For example, the passenger air transport case has a relatively high observed severity, but only a small share of passenger activity is estimated to be exposed to chatbot interactions, which keeps the overall market-wide estimate of the financial risk relatively low.

AI failure category	Case-study sector(s)	Indicative annual financial risk (% of global market)
Biased or discriminatory AI-powered decision-making	Tenant screening; student loan underwriting; online tutoring	0.3%
Hallucinations and fabricated outputs	Management consulting; litigation legal services	2.5%
Customer-facing chatbot failure	Passenger air transport	0.1%
Agentic AI failure	Enterprise use of agentic AI across sectors	3.6%

Table 1: Headline estimates of annual financial risk by AI failure category



* Market revenue refers to the total annual value of economic activity in a market, measured by how much money is spent or earned each year. Alma researchers used this figure to estimate the potential financial impact of AI failure.



These results should be interpreted as an indicative estimate of the potential scale of financial risk based on the cases examined, rather than a precise point estimate in recognition of the uncertainty associated with the analysis.

The findings should be considered in the context of the nascency of the evidence base, given the recent and rapidly evolving nature of AI, its adoption and associated implementation challenges. The analysis also relies on evidence available in the public domain, which introduces further uncertainty.

The seven case studies provide useful real-world examples but are not representative of all AI failures: many incidents are difficult to monetise, are not publicly disclosed, or are not widely reported.

Conclusion

The findings illustrate that AI failure can create material financial exposure across a range of business applications, reinforcing the importance of effective AI governance and risk management as adoption continues to grow.

The results provide a snapshot based on the evidence available today and the current state of AI technology. As AI adoption expands, more organizations, sectors, and business processes are likely to become exposed to AI-related failures. At the same time, the technology itself is evolving rapidly: improvements in model performance, testing and governance may reduce some existing risks, while new use cases and more autonomous forms of AI may introduce different types of failure or increase exposure to existing ones. The scale and nature of financial risk may therefore change over time, and the estimates should not be interpreted as forecasts of future AI failure costs.

2.1 Introduction and research objectives

AI is increasingly being integrated into business processes, decision-making and customer-facing activities. Alongside the potential benefits, its use can also give rise to failures that result in financial, operational, legal and reputational consequences.

Against this backdrop, BSI commissioned Alma Economics to explore the potential economic costs associated with AI failure. The research sits within BSI's wider work on AI governance and aims to provide an evidence-based indication of the scale of risk AI failure may create for businesses and wider markets. We also conducted opinion polling of 1,287 business leaders across nine markets and seven sectors.

Before many organizations deploy AI outside of controlled environments, it is valuable to explore what is at stake if AI fails. Beyond the hype, we want to bring tangible context to the importance of effective AI governance and risk management.

The research seeks to understand the impact different types of AI failure can have on individual organizations, as well as the potential scale of these impacts across industries, and at national or global level. By considering different failure types, the analysis also aims to illustrate how the scale and nature of financial exposure can vary depending on how AI is used and the context in which a failure occurs.

To undertake this analysis, we identified a shortlist of real-world examples of AI failure, then developed a consistent framework for quantifying and scaling their financial impacts. We used the resulting estimates to assess the potential scale of financial exposure across different types of AI failure.

The intention is not to estimate a single overall cost of AI failure, but to provide indicative evidence on the order of magnitude of financial risk associated with different failure types.

The findings are intended to support broader understanding of the potential economic significance of AI failure and to inform discussion around the importance of effective AI governance and risk management.

We believe it's imperative to understand the financial risk from AI failures. Unmanaged, this risk could have long-term effects on economic performance and undermine the projected productivity and efficiency gains markets are hoping AI will produce.



2.2 Definition and scope of AI failure

For the purposes of this research, AI failure is defined as a situation in which an AI system makes an error, malfunctions, degrades in performance, or produces a biased output, where this behaviour is generally unexpected by the developer, deployer and user.

This definition draws on the [taxonomy developed by the Center for Security and Emerging Technology \(CSET\)](#), which identifies six mechanisms through which use of AI can lead to harm, grouped into intentional and unintentional harm.



Within this context, AI failure is classified as a form of unintentional harm and is the most frequently observed harm mechanism in the [AI Incident Database](#) examined by CSET.

This distinction is important for defining the scope of the analysis. Not all harmful outcomes involving AI constitute an AI failure. For example, harm may arise through the deliberate design or misuse of an AI system, attacks on AI systems, failures of human oversight, or the consequences of deploying an otherwise functioning system in an unsuitable context.

The research therefore focuses specifically on cases where the behaviour or output of the AI system itself represents an unexpected error, malfunction, performance failure or bias. This provides a consistent basis for identifying and comparing cases while distinguishing AI failure from the wider range of mechanisms through which AI can cause harm.

2.3 Structure of the report

- Section 3 explores the survey findings
- Section 4 sets out the methodology, including identification and selection of case studies, the modelling framework used to estimate financial risk, the approach to scaling estimates across geographies, and the sensitivity analysis.
- Section 5 presents the main results, discusses their interpretation and limitations, and considers the potential relevance of the failure types examined to a wider range of sectors and applications.
- Appendix I provides a summary of the modelling inputs used for each case study.
- Appendix II presents additional results, including case-study and regional breakdowns and the sensitivity analysis.

2.3 Key findings

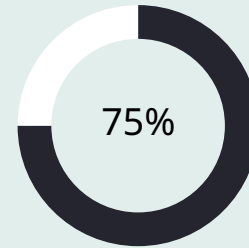
Global market revenue at risk

AI failure could put approximately 0.1% to 3.6% of relevant global market revenue at risk each year.

3.6%

Failure is not a rarity

Three quarters of organizations have experienced an AI failure in the last year, with 92% experiencing more than one.



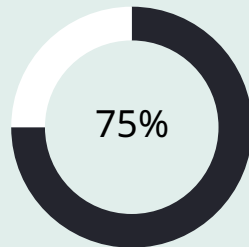
Governance is inconsistent

Nearly half have a process for managing AI incidents (up from 29% last year). However, there is an AI governance confidence-action gap that needs closing.

47%

Confidence is undeterred

Three quarters say AI projects deliver ROI and 81% say AI has delivered tangible benefits. Confidence is high and most report positive returns.



Investment is increasing

Nearly eight in ten expect AI investment to increase, even as 68% report costs have already risen and 71% expect costs to rise further.

78%

Risk can be managed

Much like in cybersecurity where good governance and risk management helps reduce the impact of incidents, AI governance can offer similar gains for managing failure and reducing the potential financial risk presented.

2.4 Considerations and limitations

The methodology used for the modelling in this report is robust, but it is important to make clear limitations to the findings. AI is only just starting to be deployed at scale across businesses and society, and the transition to an AI-enabled world is far from complete. Thus, this research should be treated as an indicative snapshot based on currently available information, rather than a direct forecast or precise quantification.



Focus on economic rather than wider costs

The modelling only looks at the economic impacts of failures within the market in which they occur (for example the passenger air transport market). This does not account for broader ramifications for individual businesses, such as reputational cost, the long tail of a trust breakdown, or the internal costs associated with responding to a crisis. Indeed, to qualify for the modelling, failure case studies had to have had visible financial consequences, necessarily excluding incidents for which the main consequences have been reputational, operational, productivity-related or otherwise difficult to quantify. As a result, the study could understate total economic harm because indirect costs are excluded.

Limited case study pool

The evidence we have drawn on comes from disclosed incidents with measurable impacts. We recognize that organizations may have limited

incentive to disclose failures, so the full picture of AI failure is not available to be studied.

We consider seven incidents across four failure categories. While a relevant sample to extrapolate from, this remains a limited evidence base from which to generate market-level estimates. The sample is inevitably also biased towards publicly disclosed incidents with measurable impacts, which may disproportionately represent larger, more severe or legally consequential events. Therefore, the figures should be treated as illustrative estimates rather than exact estimates of the typical cost of AI failure.

In addition, the calculation takes the financial severity observed in an individual case and combines it with market size, AI adoption/exposure and failure likelihood. We recognize that the severity (or lack of consequence) of one incident may not be representative of what a typical failure would cost.

Meaning of market revenue

The impact from different use cases is calculated in slightly different ways. The categories use different sectors, different severity denominators and some differing modelling inputs.

This means that when we talk about market revenue, we are referring to the total amount of money generated or spent each year in the market or activity relevant to each case study.

In most cases, this is the annual revenue of the relevant sector or market. For example, the chatbot case uses annual passenger air transport revenue. In this use case AI failure could have significant economic consequences for businesses across similar sectors if such vulnerabilities are not detected and rectified. For agentic AI, the relevant measure is instead worldwide spending by organizations on foundation generative AI models.

The headline estimates therefore relate to the specific market or activity considered in each case.

With these limitations in mind, we are confident that the research provides an estimate of the potential ramifications of AI failure, showing that failure at scale could have sizeable consequences for individual businesses, the markets they sit in, and the wider AI industry. The fact our survey finds that 75% of businesses have experienced some form of AI failure in the last year underscores that failure is a real and present challenge, one that will undoubtedly have a cost.

The results are a snapshot of an exceptionally fast-moving technology. The estimates depend on adoption, technology, evidence and governance environment. Better models or governance could reduce risk, while broader deployment could increase it. Additional regulatory regimes could also increase the financial consequences of certain failures.

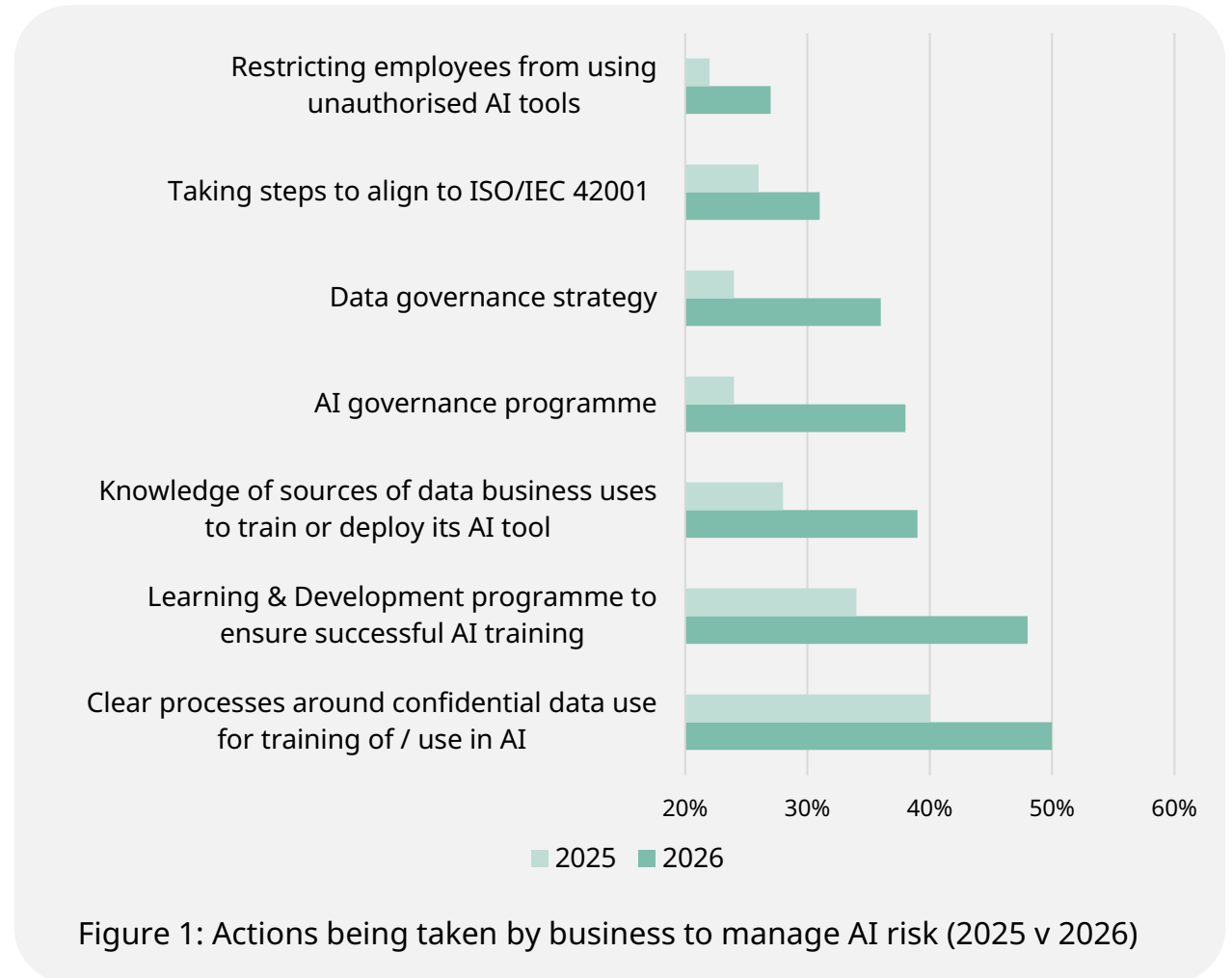
The market revenue cost figure cited in this report is not a precise expected loss. Instead, it is an indicative estimate of the potential scale of financial exposure to the relevant market in which the case studies are based (e.g. passenger air transport). Ultimately, it serves as a reminder of the critical need for AI governance for individual organizations and guardrails to manage agentic AI.

3.1 State of AI governance

Surveying more than 1,250 business leaders globally*, it's clear that AI governance is improving year-on-year. Yet given the pace at which AI is advancing, the question remains as to whether this is fast enough.

Organizations are looking to confidently keep pace with developments, particularly with AI agents coming to the fore. Yet less than two fifths (38%) say their organization has established an AI governance programme. Again, while this is up from 24%, this remains some way off supporting widespread responsible AI deployment.

Positively, a growing proportion appear to be recognizing the value of skills development, with half (50%) saying their organization has clear processes around use of confidential data for training of or use in AI, up from 40%. Almost half (48%) now have programmes to ensure successful delivery of AI training. Organizations are also getting better at assessing AI risk, with 49% now having a risk assessment approach (up from 30%).



3.2 Business leader perspectives

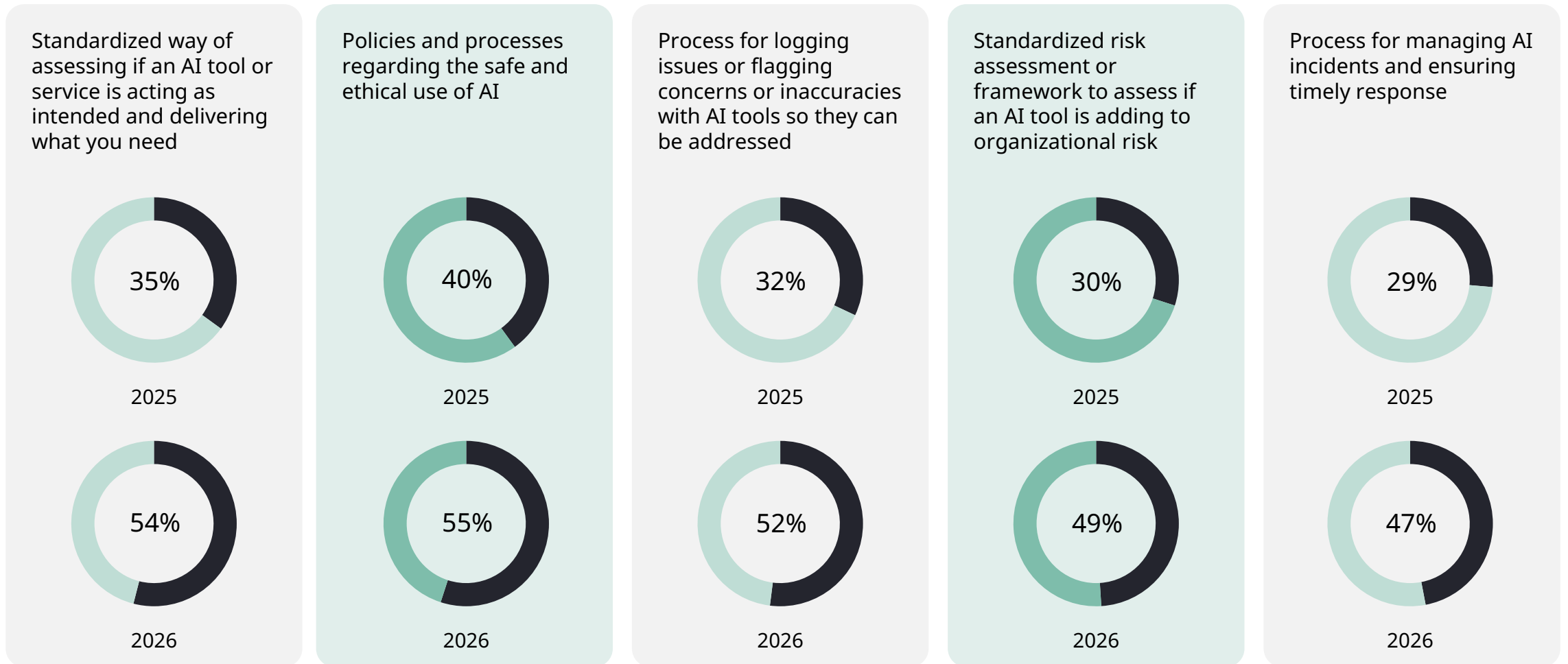


Figure 2: Processes the organization has in place (2025 v 2026)

3.3 The confidence-action gap

Last year BSI identified a significant governance gap, opening businesses up to risk around AI. This governance gap remains and extends to other areas, with a mismatch between actions organizations are taking and confidence leaders have in the organization's ability to respond.

Data and transparency

There is a clear gap between governance and workforce awareness: while 68% of organizations say they are transparent about AI data sources and algorithms, only 39% of employees understand what data is being used to train or deploy AI tools, suggesting transparency initiatives are not yet translating into organizational understanding and confidence.

Governance and standards

Similarly, organizations appear to be prioritizing AI policy implementation ahead of formal governance maturity, with two thirds adopting AI policies but only 38% having established an AI governance programme and 31% aligning to the ISO/IEC 42001 standard.

Training

When it comes to building employee skills in AI literacy and use, the data shows that many organizations are indeed investing in AI skills training, with 69% teaching employees how to use AI outputs critically. Yet fewer than half (48%) have a dedicated learning and development programme in place to support AI adoption at scale.

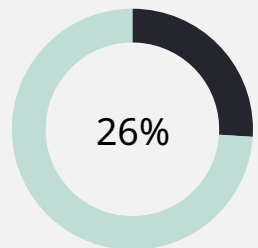


3.4 AI failure – a clear and present challenge

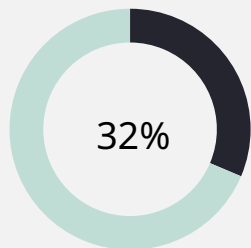
75%

Have experienced an AI failure in the last year.

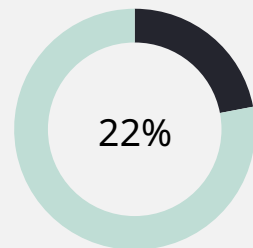
AI failure is not theoretical. In fact, three in four leaders said their company had encountered an AI failure in the preceding 12 months, rising to 88% of those who were developing their own AI tools or products, with notable costs. AI failures are being encountered more commonly in large businesses (83% versus 57% of SMEs), and skew towards the technology (82%), financial services (82%) and medical device (94%) sectors. Nearly one in ten leaders said their company has experienced up to 100 failures in the last year, with 92% experiencing more than one incident.



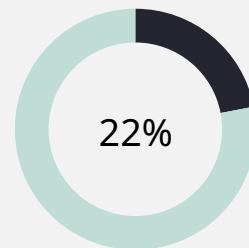
During experimentation / pilots



With an internal tool



With an external tool



With both internal and external tools

Figure 3: Types of failures experienced by businesses in the year to August 2026



Figure 4: Impact of AI failures

3.5 Agentic AI use is gathering speed

Organizations are using AI in different ways, but overwhelmingly they are using agentic AI. This refers to a collection of AI agents that interact autonomously. The agents can act, make decisions and complete a task with minimal human intervention. All but 10% are using it in some form.

As AI systems and agents gain greater autonomy, the concern is that failures may become more difficult to detect and potentially more costly.

90%

Businesses using Agentic AI in some form

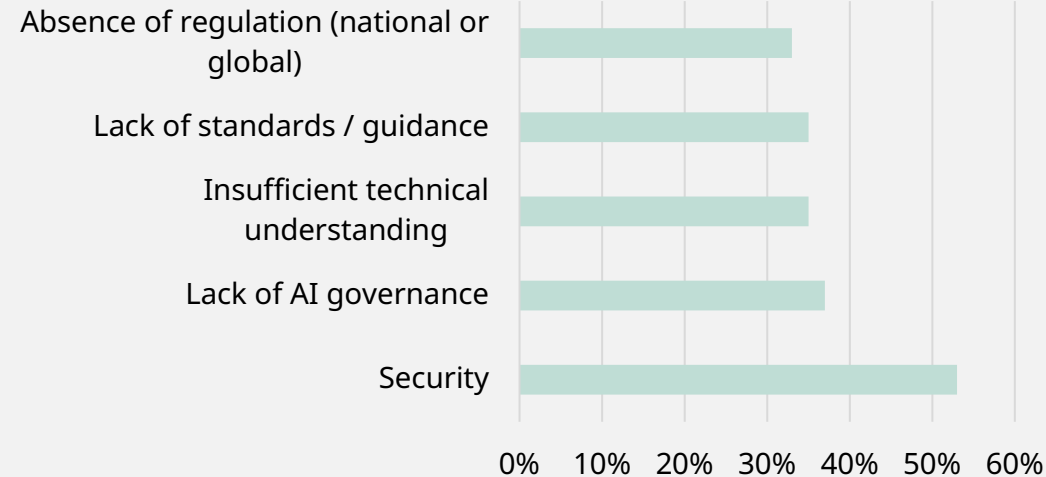


Figure 5: Concerns within the organization around using Agentic AI

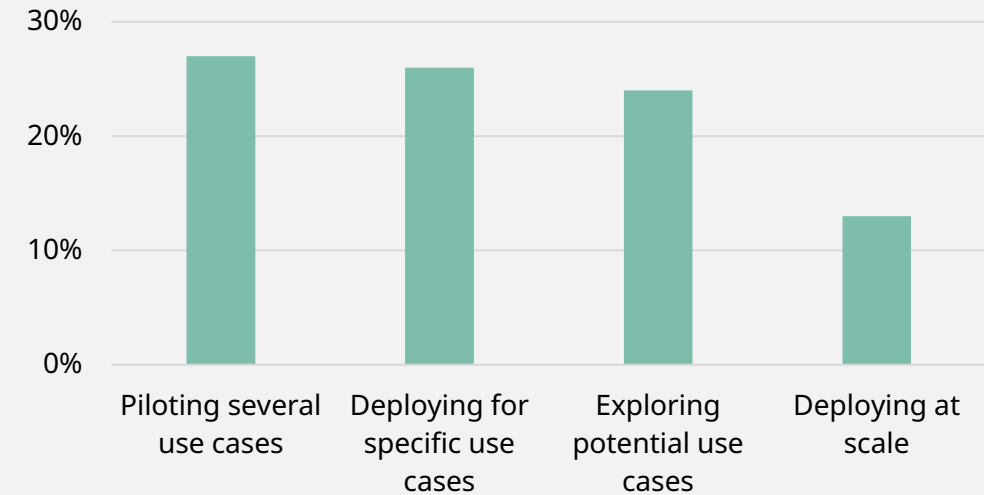


Figure 6: Type of Agentic AI use

4. Economic model: Methodology

4.1 Identification and selection of case studies

AI failure and the resulting impact will depend on the precise application of AI and the context within which it is used. It would therefore be impossible to model AI failure through a systematic modelling approach which properly captures the range of impacts and uses.

A single source of evidence on AI failure that can be used for economic analysis currently does not exist. Given these challenges, a case study approach was used as the main structure of the analysis. This utilizes existing real-world examples to anchor an understanding of AI failure and its severity. This allows for the extrapolation to an industry-wide level.

4.1.1 Case study identification

A broad desk-based search was undertaken to identify potential real-world cases of AI failure suitable for the modelling. The search drew on a range of sources, including:

- Public repositories maintained by third-party sources, including the [OECD AI Incidents and Hazards Monitor \(AIM\)](#), the [MIT AI Incident Tracker](#), and the [AI Incident Database](#).
- News and media reporting.
- Industry and consultancy reports (e.g., [EY](#) and [McKinsey](#) reports).
- Targeted research using general search engines.

For each potentially relevant case, we recorded a summary of the incident, the organization involved and wider industry, and relevant source material.

We bolstered this with available information on the financial or otherwise quantifiable consequences of the failure. Cases were then assessed against the selection criteria to determine whether they were suitable for inclusion in the modelling.



4.1.2 Case study selection criteria

Identified cases were assessed against three broad criteria:

1. The incident had to fall within the definition of AI failure set out in Section 2.2. This meant that the case needed to involve an unexpected error, malfunction, deterioration in performance or biased output from an AI-enabled system. Incidents were excluded where the harm resulted mainly from other causes, such as deliberate misuse, attacks on AI systems or failures of human oversight, rather than from a failure of the AI system itself.

The same applied to issues relating primarily to the development or governance of AI, such as the unauthorised or inappropriate use of data for model training, where there was no identifiable failure in the behaviour or output of the AI system itself.

2. The case needed to provide a quantifiable measure of the impact associated with the failure. As described in Section 4.2, the modelling uses the observed impact of each case to derive a measure of failure severity, expressed relative to an appropriate reference value such as company revenue, transaction value or project value. Relevant impacts could include, for example, fines or court sanctions, legal settlements, compensation or refunds paid to customers, and other identifiable financial losses directly associated with the failure.

A case therefore needed to include a reported financial cost or another sufficiently well-defined quantitative impact that could be translated into a financial measure. In some cases, these financial consequences may also reflect harms experienced by consumers or other affected individuals, for example, where compensation or settlement payments are made in response to discriminatory treatment or incorrect information.

However, the modelling captures these harms only to the extent that they result in a measurable financial consequence. Cases where the impact was mainly reputational, operational or otherwise not quantified were excluded, as there was no clear measure to use in the modelling.

3. Failure needed to have occurred in the context of real-world deployment or operational use. Cases relating only to the development, testing or piloting of an AI system were excluded where the technology was not subsequently put into use. Although such cases can involve substantial expenditure, these costs are more appropriately interpreted as development or testing costs associated with an unsuccessful technology investment, rather than as the financial impact of an AI failure during actual use.

4.1.3 Overview of the shortlisted case studies



The seven shortlisted case studies were grouped into four categories, based on the type of AI failure involved:

- Biased or discriminatory AI-powered decision-making – cases where an AI-enabled system produces outcomes that disadvantage set groups or individuals.
- Hallucinations and fabricated outputs – cases where a generative AI system produces false, inaccurate or fabricated information that is presented as factual.
- Customer-facing chatbot failure – where an AI-enabled chatbot provides incorrect or misleading information to customers.
- Agentic AI failure – where an AI system is given a degree of autonomy to carry out tasks or actions and produces unintended outcomes, including excessive or uncontrolled costs.

For each case, the observed financial impact is expressed relative to the most relevant reference value, such as company revenue, transaction value or project value. These percentages are used as the case-specific measures of severity in the modelling described in Section 4.2.

These case studies are aimed at underpinning the analysis using real-world data, rather than singling out a specific example of AI failure or company.

Shortlisted case studies

Biased or discriminatory AI-powered decision-making

1. Tenant screening (US): A tenant-screening provider faced a class-action lawsuit alleging that its algorithmic scoring system disproportionately disadvantaged Black and Hispanic rental applicants, including those using housing vouchers. The case was settled, with the provider making a financial payment and agreeing to changes in how its screening service is used.
2. Student loan underwriting (US): A student-loan provider settled with a state authority relating to its use of AI-based underwriting models. The models were used to inform decisions on loan eligibility, terms and pricing, and were alleged to create a risk of disproportionate harm to Black and Hispanic applicants.
3. Recruitment for online tutoring (US): An online tutoring provider faced legal action over use of an automated recruitment system that allegedly discriminated against older applicants. The system applied age-based screening criteria that resulted in some applicants being automatically rejected. The case was settled.

Hallucinations and fabricated outputs

4. Management consulting (Aus): A professional services firm produced a government-commissioned report containing AI-generated errors, including fabricated quotations and references to non-existent academic sources. The errors were identified after publication, leading to the report being revised and part of the contract fee being refunded.
5. Litigation legal services (US): Two lawyers submitted a court filing containing legal cases and citations fabricated by a generative AI tool. The lawyers and their firm were subsequently sanctioned for failing to verify the material before submitting it to the court.

Shortlisted case studies

Customer-facing chatbot failures

6. Passenger air transport (Canada): An airline customer relied on incorrect information provided by the airline's customer-service chatbot about the terms of a specific fare policy. When the airline later refused to apply the fare on the basis that the chatbot's advice did not reflect its policy, a tribunal held the airline responsible for the information provided and awarded compensation to the customer.

Agentic AI failure

7. Agentic AI in enterprise applications (Global): A major technology company used an agentic AI system for an internal data-matching task. The project generated substantially higher costs than planned, and the cost overrun went undetected for approximately about five months before the project was ultimately discontinued.



4.2 Modelling approach

Each of the shortlisted case studies provides an observed financial impact associated with a specific AI failure. The analysis uses these observed impacts as a starting point to estimate the potential revenue at risk from similar failures across the wider market in which the failure occurred, using a combination of the four key elements in the figure below.

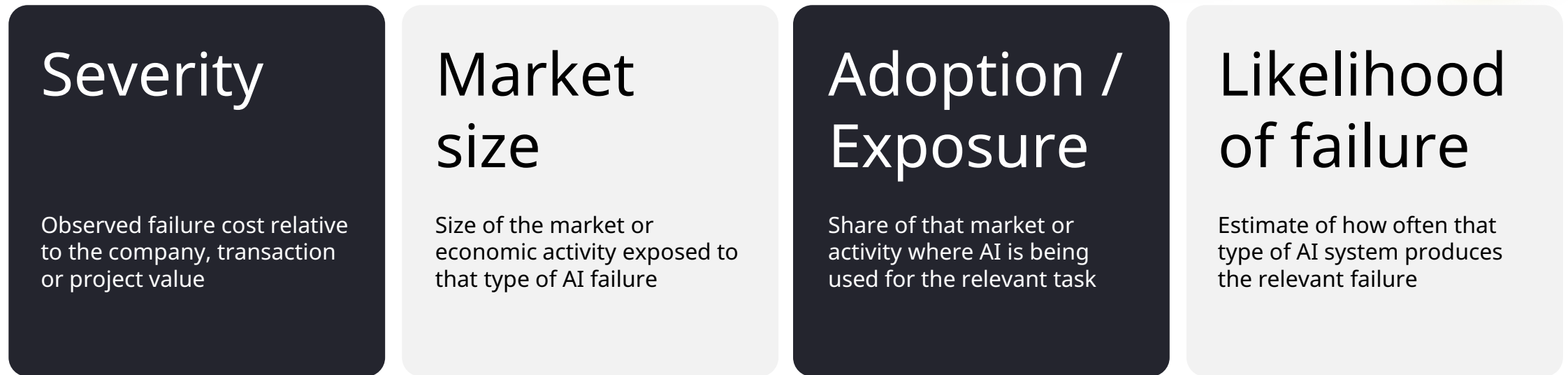


Figure 7: Key elements of the modelling framework



For each case study, the observed financial impact is first expressed relative to the most relevant reference value, such as annual revenue, the value of a transaction or the value of a project. This provides a measure of the failure's severity. The model then combines this with the size of the relevant market, the extent to which AI is used for the relevant activity, and an estimate of how often that type of failure occurs.

Together, these four elements are used to estimate the annual financial value potentially at risk within the relevant sector and geography. In broad terms, the calculation can be expressed as:

Estimated annual revenue at risk = Severity × Market size × Adoption/exposure × Likelihood of failure

The precise definition and data used for each element vary across the case studies, reflecting differences in the type of AI failure and the evidence available. The four elements of the modelling framework are described on the next page, highlighting the common logic as well as key differences in how they are applied. A summary of the measures used for each element in each case study is provided in Appendix I.

4.3 The modelling framework

4.3.1 Severity

Severity captures the scale of the financial impact when a failure occurs. This element converts the financial impact observed in each case study into a relative measure that can scale the impact beyond the individual incident. This is done by expressing the financial cost as a share of the most relevant reference value for the activity affected, which can then be extrapolated to an industry level.

In most cases, the observed cost is expressed as a share of the organization's estimated annual revenue. Where the failure relates to a specific contract, transaction or project, that value provides a more meaningful basis (for example, for the management consulting case or the passenger air transport case).

The underlying logic is therefore consistent across the case studies, while the reference value is selected to best reflect the scale of the activity directly affected by the failure.

4.3.2 Market size

The identified severity is then scaled up to the industry level, essentially reflecting the value of impact should this failure be observed at the same level across the sector. The market size element represents the scale of the economic activity potentially exposed to the type of AI failure considered in each case study. In most cases, this is measured using the annual revenue of the relevant industry or market in the country where the failure occurred. The market is selected to reflect as closely as possible the activity in which the AI system was being used. For example, the passenger air transport case uses passenger revenue generated by Canadian air carriers.

The agentic AI case is treated differently because the failure is linked to the use of agentic AI across organizations rather than to a particular industry. The relevant economic base is therefore worldwide spending on foundation AI models, rather than the revenue

of the technology sector of the original case, providing a measure more closely aligned with the activity exposed to this type of failure.

4.3.3 Adoption/exposure

AI adoption and exposure determine how much of the relevant market is potentially exposed. The model therefore adjusts for the share of the market in which AI is being used for the activity associated with the failure.

AI adoption varies. Estimates are generally drawn from industry reports and survey evidence and relate as closely as possible to the specific AI use and geography of the case study. For example, the legal services case study uses the share of US legal professionals using generative AI, while the student loan case uses the share of US financial institutions using AI to support credit underwriting and decision-making. Where country-specific evidence is not available, broader regional or global estimates are used.



For the passenger air transport case, adoption and exposure are treated as two separate elements. The first captures the share of airlines using AI for customer service, while the second estimates the share of passengers who interact with an airline chatbot. These two measures are combined to estimate the share of passenger activity that is exposed to this specific type of AI failure.

4.4 Likelihood of failure

The model also adjusts for the likelihood of failure, which may depend on the AI technology involved and the context in which it is being used. This element represents how often the relevant type of AI failure is expected to occur among the organizations, users or transactions exposed to that AI use. Where possible, this is based on evidence of the frequency of the specific failure being modelled, drawn from academic studies, industry reports or other relevant sources.

The precise measure varies across case studies. For example, for hallucinations and fabricated outputs, the likelihood measure is based on the share of GPT-4o-generated citations found to be fabricated in the most relevant and up-to-date academic study. For the passenger air transport case, it is based on the estimated probability that an AI customer-service interaction provides incorrect policy information, drawing on evidence from a specific airline case.

A separate likelihood-of-failure parameter is not applied to the three case studies involving biased or discriminatory AI-powered decision-making. In these, the failure is treated as a recurring pattern in the way the system makes decisions, rather than as a discrete event. The observed bias or discriminatory effect is therefore treated as part of the underlying pattern of outcomes, without applying an additional probability of failure.

4.5 Scaling estimates across geographies

To estimate the potential market revenue at risk at the global level, the models are scaled beyond the country in which the original failure occurred. The global scale-up follows the same overall framework in Section 4.2 but adjusts the expected inputs to vary across geographies.

Two of the four modelling elements are kept unchanged. Severity remains anchored in the financial impact observed in the original case study, on the basis that this provides the empirical measure of the scale of the failure. Likelihood of failure is also held constant across geographies. The two elements adjusted geographically are market size and adoption/exposure.

For market size, the relevant industry or economic activity identified in the case-study model is replaced with the corresponding market value for five broad world regions: North America, Europe, Asia-Pacific, Latin America, and the Middle East and Africa.

Where suitable data is available, the model is estimated separately for each region, and the regional estimates are then summed to produce a global estimate.

The adoption/exposure rate is adjusted geographically where the original case-study model relies on a country-specific estimate. Applying the adoption rate uniformly to all regions would, for example, imply the same level of AI use in the US and Latin America. Instead, the modelling uses the IMF AI Preparedness Index (AIPI)¹ as a proxy. Where the adoption/exposure estimate used in the original case study is already global, no adjustment is applied, and the original rate is retained across regions.

The AIPI provides comparable scores across 174 countries and assesses preparedness for AI adoption across four areas: digital infrastructure; human capital and labour-market policies; innovation and economic integration; and regulation and ethics.

It is not a direct measure of AI adoption, but its broad country coverage and focus on the conditions that support adoption make it a useful basis for comparing relative levels of AI preparedness across geographies. For each region, the adoption/exposure rate from the original case study is adjusted based on the relative difference between the case-study country's AIPI score and the corresponding regional score. In simplified terms:

$$\text{Regional adoption/exposure rate} = \text{Case-study adoption/exposure rate} \times (\text{Regional AIPI score} \div \text{Case-study country AIPI score})$$

This preserves the adoption/exposure evidence identified for the original case study while allowing the rate to increase or decrease across regions in line with AI preparedness. The resulting regional rates are combined with the corresponding regional market sizes, while retaining the original severity and likelihood-of-failure assumptions, to produce the regional and global estimates of financial risk.



1. Regional AIPI scores for North America, Europe, and Asia and the Pacific are taken directly from the IMF AIPI source. Equivalent regional scores are not reported for Latin America or the Middle East and Africa (MEA). For these regions, population-weighted averages were calculated using country-level AIPI scores together with World Bank data on country populations and regional classifications. Countries with no AIPI score available were excluded.

4.6 Sensitivity analysis

The modelling relies on several assumptions and estimates, so the resulting estimates are subject to uncertainty. Sensitivity analysis was included to illustrate the impact of this uncertainty on the results, namely in relation to two key model parameters, severity and likelihood of failure, which are varied.

For each of these parameters, three settings are available:

1. A base case, which uses the central estimate described in Section 4.2;
2. A low case, which applies a multiplier of 0.8; and
3. A high case, which applies a multiplier of 1.2.

The sensitivity analysis focuses on these points because they capture two central dimensions of the risk being modelled: the financial impact of a failure and the frequency with which it occurs. Market size and adoption/exposure remain at their base-case values throughout. The sensitivity analysis is intended to illustrate how responsive the results are to variation in these assumptions, rather than to provide confidence intervals or a full range of possible outcomes.



5. Findings and interpretation

5.1 Main results

The modelling produces indicative estimates of the annual financial risk associated with the four types of AI failure examined. Across the case studies examined, the modelling suggests that AI failure could cost approximately 0.1% to 3.6% of market revenue each year, depending on the type of failure and the sectors or activities considered, as shown in Table 1.

The financial impacts used to anchor the modelling take different forms across the case studies. These include legal settlements and court sanctions, compensation paid to affected customers, refunds or revenue returned following an AI failure, and excess expenditure caused by the use of an AI system. More broadly, these costs provide observable measures of the financial consequences borne by organizations as a result of AI failure.

In some cases, particularly settlements and compensation payments, they may also indirectly reflect harm experienced by consumers or other affected individuals, although they should not be interpreted as a comprehensive measure of these wider harms.

AI failure category	Case-study sector(s)	Indicative annual financial risk (% of global market)
Biased or discriminatory AI-powered decision-making	Tenant screening; student loan underwriting; online tutoring services	0.3%
Hallucinations and fabricated outputs	Management consulting; litigation legal services	2.5%
Customer-facing chatbot failure	Passenger air transport	0.1%
Agentic AI failure	Enterprise use of agentic AI across sectors	3.6%

Table 1: Headline estimates of annual financial risk by AI failure category

The highest indicative estimate is associated with agentic AI failure, at approximately 3.6% of the relevant global market revenue, followed by hallucinations and fabricated outputs, at approximately 2.5% of the relevant global market.

Lower estimates are produced for biased or discriminatory AI-powered decision-making and customer-facing chatbot failures, at approximately 0.3% and 0.1%, respectively.

These estimates should be interpreted as the potential order of magnitude of financial exposure associated with the specific failure types and case studies examined, rather than as precise point estimates. Differences between the estimates reflect not only type of AI failure, but also the specific sectors, use cases and modelling assumptions underpinning each case study. The results should not be interpreted as a ranking of the relative risk posed by different types of AI failure.

Additional breakdowns of the results, including case-study and regional estimates, together with the outcomes of the sensitivity analysis, are presented in Appendix II.



5.2 Interpretation and limitations

5.2.1 Interpreting the headline estimates

These results should be interpreted as an indicative estimate of the potential scale of financial risk based on the cases examined, rather than as a precise point estimate, given the uncertainty associated with the analysis.

As described in Section 4.2, the estimated financial risk depends on the severity of the observed failure, the extent of relevant AI adoption within the market, and, where applicable, the likelihood of that failure occurring. The estimates therefore reflect the specific circumstances of the case studies and sectors modelled, rather than the type of AI failure in isolation.

This can be illustrated by comparing the management consulting and air transport case studies. In the latter, the observed financial impact was relatively severe, amounting to approximately 50% of the affected customer transaction.

However, the model assumes only a relatively small share of passengers interact with an airline chatbot, and that incorrect information is provided in an even smaller proportion. These factors substantially reduce the share of the overall market exposed to the failure.

By contrast, the management consulting case has a lower observed severity, at around 22% of the affected contract value, but a much larger share of the relevant activity is estimated to involve generative AI, alongside a higher assumed likelihood of relevant failure. The resulting estimate is therefore substantially higher, at around 3.5% of the global management consulting market.

This illustrates why the results should not be interpreted as evidence that one type of AI failure is inherently more costly or more serious. Instead, the estimates are shaped by the characteristics of the case studies, how AI is used within the sector, and the evidence available to inform the modelling.

5.2.2 Limitations of the evidence and modelling approach

The results should be interpreted in light of limitations in both the available evidence on AI failure and the modelling approach used to scale the observed impacts.

The results should be considered in the context of the nascency of the evidence base, given the recent and rapidly evolving nature of AI and associated implementation challenges. The case studies provide useful real-world anchors for the analysis but should not be considered representative of the full range of AI failures.

Publicly available evidence is likely to capture only a fraction of incidents. Many AI failures are inherently difficult to monetize, particularly where the consequences relate to productivity, service quality, reputation, customers or staff without a clearly identifiable financial cost. Even when organizations can quantify financial impacts internally, they may have limited incentives to disclose them publicly.

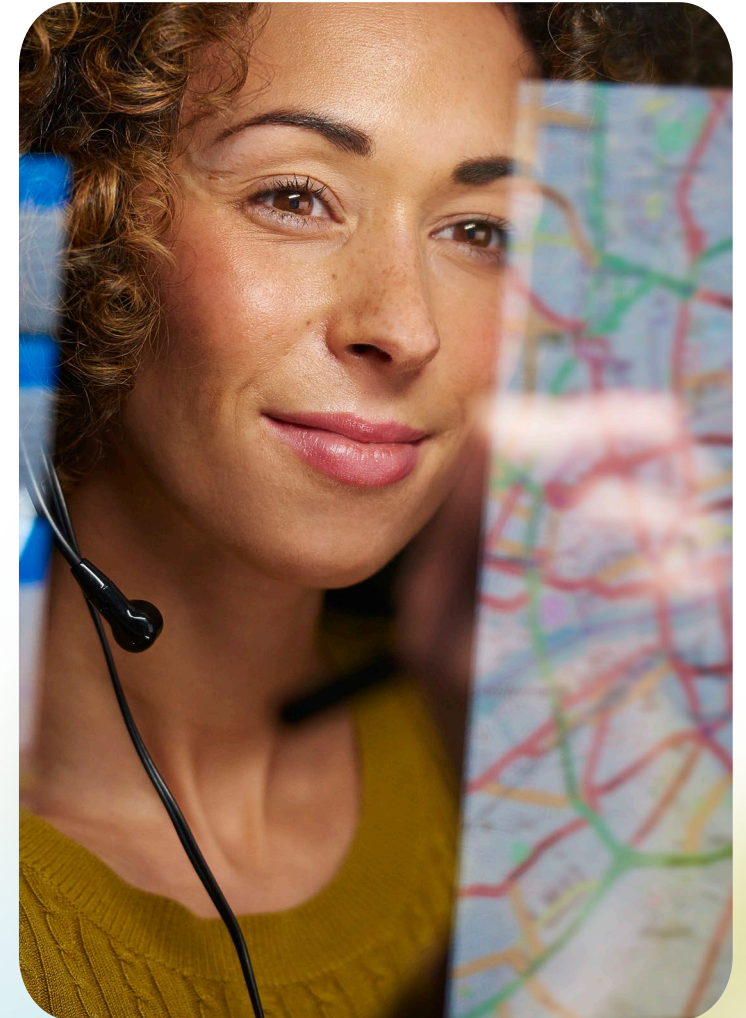
In addition, incidents with measurable impacts may never be widely reported. The available evidence is therefore selective and may disproportionately reflect larger, more severe or legally consequential incidents. This is particularly relevant for this research, as cases could only be included in the modelling where a sufficiently clear monetary or quantifiable impact could be identified.

There are also limitations associated with scaling individual case-study impacts to wider market-level estimates. Key model inputs, including market size, AI adoption/exposure and likelihood of failure, are subject to uncertainty, and different sources can provide materially different estimates of the same parameter. In some cases, direct measures of the relevant input are not available, and the modelling therefore relies on the closest suitable proxy, introducing an additional layer of uncertainty. Severity estimates are also anchored in the financial consequences of individual observed incidents.

These consequences are shaped by the specific circumstances of each case and may differ substantially across organizations, transactions or projects, limiting the extent to which they can be generalized.

The global scale-up requires further simplifying assumptions. In particular, geographic differences are approximated using the AIPI, rather than a direct measure of adoption, while the likelihood of failure is assumed to remain constant. These assumptions provide a consistent basis for extending the estimates across regions, but they inevitably simplify differences in how AI is adopted and used.

Finally, the estimates represent a snapshot based on currently available evidence and the present state of AI technology. For example, future regulatory sanctions, including those under the EU AI Act, could add to the financial consequences of certain failures, but are not included in the current estimates where they were not part of the observed cases.





However, experience from cybersecurity provides a useful comparison for how technology-related financial risks can evolve over time. Long-run evidence indicates that both the likelihood and financial impact of cyber incidents have increased as digital technologies have become more embedded in business activity.

For example, the [Cyentia Institute's 2025 Information Risk Insights Study](#), which analyzed cyber incidents between 2008 and 2024, estimated that the annual probability of a typical organization experiencing a significant cyber incident increased from around 2.5% to 9.3%, and median losses relative to the affected organization's annual revenue increased from 0.08% to 0.65%.

While AI failure differs materially from cyber risk and the same trajectory cannot be assumed, this experience illustrates how financial exposure associated with emerging technologies can evolve over time, particularly

where adoption and exposure grow faster than the governance and controls to manage them.

Therefore, as AI adoption continues to expand rapidly across organizations, sectors and business processes, with the technology itself also evolving quickly, both the frequency and nature of AI failure may alter over time.

Improvements in model performance and governance may reduce some risks, while new use cases and more autonomous forms of AI may create different types of failure or increase exposure to existing ones. The results should therefore be interpreted as indicative estimates, rather than as precise forecasts of future AI failure risk.

5.3 Wider sector implications

The case studies in this research relate to specific sectors and applications, but the underlying failure mechanisms are not unique to them. To consider where similar risks may arise, we identify characteristics of AI uses in specific sectors that may increase exposure to each failure type, drawing on the case-studies and wider secondary research. It then considers other sectors and applications that share similar characteristics.

This mapping is illustrative rather than exhaustive and does not provide a quantitative assessment of risk in the additional sectors. Rather, it highlights sectors and applications with similar conditions, where the failure mechanisms examined in this research may therefore also be relevant.

Taken together, the mapping suggests that exposure to AI failure is shaped not only by sector, but by how AI is used within this. Similar failure mechanisms may arise wherever organizations use AI for high-volume decision-making, information-intensive professional outputs, customer-facing interactions or increasingly autonomous workflows. Several sectors, including financial services, healthcare, retail, transport and technology, may therefore be exposed to more than one of the failure mechanisms considered.



AI failure type	Case-study sectors / applications	Key industry characteristics	Potentially relevant sectors / applications
Biased decision-making	Tenant-screening Student loan underwriting Recruitment	• High volume of screening • Repetitive with constrained range of relevant information (e.g. income, educational status) • Relevant characteristics correlated with socio-economic traits (e.g. income correlated with race and gender)	• Student admission applications • Medical applications for diagnostics, e.g. skin cancer • Insurance • Public sector benefits administration • Adaptive/surge pricing • Other HR tasks, e.g. performance reviews
Hallucinations and fabricated outputs	Management consulting Legal services	• Professional services • Reputational risk – providers hired for expertise/credibility • Information-dense outputs involving knowledge synthesis • Verifying each statement/source is highly effort-intensive	• Academia and research • Healthcare services • Financial services, such as accounting, audit and investment
Customer-facing chatbot failure	Passenger air transport	• High volume of customer interaction • Focus on customer experience and customised services • Need for out-of-hours support as with airlines • Significant reputational risks to revenue	• All retail services • Financial services e.g. online banking • Other transport and logistics
Agentic AI failure	Technology	• High labour costs encouraging automation • Concerns around privacy/data access • Regulatory / ethical concerns around AI decision-making	• Banking and financial services • Insurance • Customer services (e.g. independently issuing refunds)

Table 2: Illustrative mapping of AI failure types to potentially relevant sectors and applications. (Sources: McKinsey, Deloitte, Harvard Medical School)

6. Conclusion

6.1 Conclusion and key insights

The findings illustrate that AI failure can create material financial exposure across a range of business applications, reinforcing the importance of effective AI governance and risk management as adoption continues to grow.

The results provide a snapshot based on the evidence available today and the current state of AI technology. As AI adoption expands, more organizations, sectors, and business processes are likely to become exposed to AI-related failures. At the same time, the technology itself is evolving rapidly: improvements in model performance, testing and governance may reduce some existing risks, while new use cases and more autonomous forms of AI may introduce different types of failure or increase exposure to existing ones. The scale and nature of financial risk may therefore change over time, and the estimates should not be interpreted as forecasts of future AI failure costs.

The critical takeaway for leaders is to take a resilience mindset. As AI adoption gathers pace, leaders will need to be mindful of emerging risks, identify their governance gaps, and consider what specific action is required.



6.2 Actions for organizations

Horizon scanning

Good AI governance means organizations are looking ahead at what's on the horizon and reflecting this in their guardrail and governance choices.

ISO/IEC 42001 can help.

Monitor regulation

While AI regulation remains fragmented, developments continue to evolve globally. It's critical to be aware of and prepare your obligations, especially as financial implications emerge, such as the EU AI Act fines now imposed.

AI regulatory assessments can help.

Be transparent

Not just with the AI tools you choose to adopt but over the prompts and parameters you bake into them. Having this clearly understood and validated will help with safe adoption and explainability requirements.

AI performance assessments can help.

Take control

By embedding strong risk management and cybersecurity approaches as part of your AI governance commitments, you can take control of potential risks and contain them proportionately to your business.

AI foundation framework can help.

Build capability

Two organizations can have the same AI tool and get completely different results or be exposed to different failures. The difference won't be the technology; it'll be the people and processes around it, so take time to understand context and build appropriate capabilities.

AI training can help.

Prioritize collaboration

The pace of AI makes it challenging to stay ahead of changes and what they mean. Collaborate with others whether that's sector groups to explore similar use cases or organizations who are close to regulatory changes and standards like BSI.

Our newsletter will keep you updated.

6.3 AI governance priorities for leaders

BSI experts share top priorities for business leaders to progress AI governance as a strategic capability:

Set out a clear communication plan around your AI strategy, policy and plans for all stakeholders (internal and external).

Compare your current practices against best practice guidance such as AI management system standard ISO/IEC 42001.

Consider where and how you plan to use AI, taking into account your organization's contractual obligations, regulatory jurisdiction, and operating culture.

Establish AI objectives, policies and documented information. Determine staff training, competence, communication needs.

Determine your AI risks and select appropriate controls proportionate to your AI use, such as those in Annex A of ISO/IEC 42001, to manage them.

Establish accountability lines and conduct an AI risk impact assessment - what potential consequences could occur across the business and society if the system fails?

Verify your AI performance with internal and external audits to confirm effectiveness, gain market access and build trust with stakeholders.

Implement mechanisms to assess system efficiency through performance reviews and harm mitigation.

Commit to continual improvement so your system evolves as AI advances and the external environment changes.

Be clear on how you will monitor and measure your AI performance so you can spot when things go wrong and initiate the appropriate response.

7.1 Appendix I: Modelling inputs by case study

Modelling element	Measure/proxy used	Value
Severity	Settlement cost as % of estimated annual revenue	2.8%
Market size	Annual US tenant-screening market revenue	\$1.3bn
Adoption / Exposure	Share of tenant-screening services using AI (US)	60%
Likelihood of failure	N/A	

Table 3: Modelling inputs: Tenant screening (US)

Modelling element	Measure/proxy used	Value
Severity	Settlement cost as % of estimated annual revenue	0.5%
Market size	Annual US private student-lending market revenue	\$10bn
Adoption / Exposure	Share of financial institutions using AI for credit underwriting and decision-making (US)	0.5%
Likelihood of failure	N/A	

Table 4: Modelling inputs: Student loan underwriting (US)

Modelling element	Measure/proxy used	Value
Severity	Settlement cost as % of estimated annual revenue	0.1%
Market size	Annual US online tutoring services market revenue	\$2.3bn
Adoption / Exposure	Share of recruitment activity using AI (US)	35%
Likelihood of failure	N/A	

Table 5: Modelling inputs: Recruiting for online tutoring services (US)

Modelling element	Measure/proxy used	Value
Severity	Contract fee repaid as % of affected contract value	22.3%
Market size	Annual Aus management cons. market revenue	\$4.9bn
Adoption / Exposure	Share of management cons. staff using GenAI (global)	78%
Likelihood of failure	Share of GPT-4o-generated citations found to be fabricated	19.9%

Table 6: Modelling inputs: Management consulting (Aus)

7.1 Appendix I: Modelling inputs by case study

Modelling element	Measure/proxy used	Value
Severity	Court sanction as % of estimated revenue from a comparable case	28%
Market size	Estimated annual US litigation legal services revenue	\$104.3bn
Adoption / Exposure	Share of US legal professionals using GenAI	28%
Likelihood of failure	Share of GPT-4o-generated citations found to be fabricated	19.9%

Table 7: Modelling inputs: Legal services (US)

Modelling element	Measure/proxy used	Value
Severity	Compensation cost as % of affected passenger transaction value	50%
Market size	Annual passenger revenue of Canadian air carriers	CAD 32.4bn
Adoption / Exposure (1)	Share of airlines using AI for customer service (global)	75%
Adoption / Exposure (2)	Estimated share of passengers using an airline chatbot	4.7%
Likelihood of failure	Estimated probability that an AI customer-service interaction provides incorrect policy information	3.8%

Table 8: Modelling inputs: Passenger air transport (Canada)

Modelling element	Measure/proxy used	Value
Severity	Excess spend as % of total AI deployment cost	89.6%
Market size	Worldwide end-user spending on foundation GenAI models	\$11.4bn
Adoption / Exposure	Share of enterprise AI deployments qualifying as “true agents”	16%
Likelihood of failure	Share of AI initiatives delayed or cancelled due to unexpected costs	25%

Table 9: Modelling inputs: Agentic AI (Global)

7.2 Appendix II: Additional results tables

Type of AI failure	Sector	Annual financial risk - Case study country	Global market size	Annual financial risk - Global	Annual financial risk (% of global market size)
Biased or discriminatory AI-powered decision-making	Tenant screening	\$22,181,250	\$5,800,000,000	\$73,186,582	1.26%
	Student loan underwriting	\$18,853,695	\$26,000,000,000	\$40,095,951	0.15%
	Online tutoring services (recruitment)	\$850,333	\$10,420,200,000	\$3,068,472	0.03%
Hallucinations and fabricated outputs	Management consulting	\$169,401,464	\$330,210,000,000	\$11,415,930,063	3.46%
	Litigation legal services	\$1,626,737,170	\$277,200,000,000	\$3,552,416,889	1.28%
Customer-facing chatbot failure	Passenger air transport	\$15,657,743	\$682,000,000,000	\$458,083,286	0.07%
Agentic AI failure	Cross-sector (agentic AI in enterprise applications)	\$409,861,667	\$11,438,000,000	\$409,861,667	3.58%

Table 10: Case-study/sector-level results

Sector(s)	Region	Global market size	Annual financial risk - Global	Annual financial risk (% of global market size)
Tenant screening; Student loan underwriting ; Online tutoring services	N. America	\$18,037,600,000	\$62,216,121	0.34%
	Europe	\$9,004,400,000	\$28,406,811	0.32%
	Asia & Pacific	\$7,588,700,000	\$12,588,654	0.17%
	Latin America	\$4,181,300,000	\$7,912,749	0.19%
	MEA	\$2,741,200,000	\$5,226,670	0.19%

Table 11: Biased or discriminatory AI-powered decision-making

Sector(s)	Region	Global market size	Annual financial risk - Global	Annual financial risk (% of global market size)
Management consulting; Litigation legal services	N. America	\$241,495,732,738	\$6,108,161,316	2.53%
	Europe	\$162,816,946,529	\$4,072,105,944	2.50%
	Asia & Pacific	\$136,229,016,051	\$3,140,928,000	2.31%
	Latin America	\$38,321,247,982	\$765,365,079	2.00%
	MEA	\$28,547,056,701	\$881,786,611	3.09%

Table 12: Hallucinations and fabricated outputs

Sector(s)	Region	Global market size	Annual financial risk - Global	Annual financial risk (% of global market size)
Passenger air transport	N. America	\$148,676,000,000	\$99,862,156	0.07%
	Europe	\$181,412,000,000	\$121,850,154	0.07%
	Asia & Pacific	\$235,290,000,000	\$158,038,734	0.07%
	Latin America	\$36,828,000,000	\$24,736,497	0.07%
	MEA	\$79,794,000,000	\$53,595,744	0.07%

Table 13: Modelling inputs: Customer-facing chatbot failure

Sector(s)	Region	Global market size	Annual financial risk - Global	Annual financial risk (% of global market size)
Cross-sector (agentic AI in enterprise applications)	N. America	\$3,633,650,350	\$130,205,804	3.58%
	Europe	\$2,548,125,874	\$91,307,844	3.58%
	Asia & Pacific	\$3,256,573,427	\$116,693,881	3.58%
	Latin America	\$617,034,965	\$22,110,420	3.58%
	MEA	\$1,382,615,385	\$49,543,718	3.58%

Table 14: Modelling inputs: Agentic AI failure

7.3 Annual financial risk (% of global market size)

Severity scenarios	
Low	0.22%
Base	0.28%
High	0.33%

Table 15: Sensitivity analysis of headline results (Biased or discriminatory AI-powered decision-making)

	Likelihood scenarios		
Severity scenarios	Low	Base	High
Low	0.04%	0.05%	0.06%
Base	0.05%	0.07%	0.08%
High	0.06%	0.08%	0.10%

Table 16: Sensitivity analysis of headline results (Hallucinations and fabricated outputs)

	Likelihood scenarios		
Severity scenarios	Low	Base	High
Low	1.58%	1.97%	2.37%
Base	1.97%	2.46%	2.96%
High	2.37%	2.96%	3.55%

Table 17: Sensitivity analysis of headline results (Customer-facing chatbot failure)

	Likelihood scenarios		
Severity scenarios	Low	Base	High
Low	2.29%	2.87%	3.44%
Base	2.87%	3.58%	4.30%
High	3.44%	4.30%	5.16%

Table 18: Sensitivity analysis of headline results (Agentic AI failure)

8.1 Your path to trusted AI

No two organizations are at the same stage of AI maturity or need the same things.

Whether you need to upskill teams, manage AI risk, comply with regulations, or verify your AI systems or governance practices, there are different paths you can take that help build stakeholder trust. We can support, whatever your priority.

Understand



Prove



Assess



Sustain



Build



8.2 How BSI can help

Understand

You know you are using AI but not always sure where or who owns it.

- [Foundation training](#)
- [AI Standards](#)
- [AI insights, research and reports](#)
- [AI literacy](#) for your whole workforce



Assess

Assess your exposure and where you invest first. You don't need to address it all at once.

- [AI foundation framework](#)
- [ISO/IEC 42001 pre-certification](#)
- [AI risk management product assessment](#)



Build

Build capability - you know your gaps, now work out how best to close them.

- [AI training](#) – four routes by role
- [ISO/IEC 42001 implementation](#)
- [BSI standards library](#)



Prove

Build confidence in your governance.

- [ISO/IEC 42001 certification](#)
- [AI Notified body](#)
- [Conformity assessments \(biometrics, Medical devices\)](#) EN 18286 certification
- [AI testing and performance](#)
- [AI Performance for LLMs](#)



Sustain

AI regulation is continually evolving, which means it stays fragmented. You need to stay ahead of the conversation and latest standards, so your governance adapts as new rules and risks arrive – that's where we come in at the one stage that never closes.



8.3 Resilience to failure

Explore how the existing suite of international standards can help you build resilience around the modelled AI failures.

AI failure category	Standards that can help
Biased or discriminatory AI-powered decision-making	ISO/IEC 42001 helps address bias management through risk assessment requirements at the organizational level, the need to document data on potential bias issues and make recommendations based on historical social biases, as well as outlining requirements on unwanted bias and controls related to fairness. Continuous monitoring requirements also help minimize bias risks that emerge. BS EN 18286:2026 Quality management system for regulatory purposes manages risk for each AI system including risk of inequality and unlawful discrimination and will provide presumption of conformity with the EU AI Act for high-risk AI systems put onto the EU market. ISO/IEC TR 24027:2021 Bias in AI systems and AI-aided decision-making covers sources of unwanted bias, measurement techniques, bias assessment for human cognitive biases, data bias and engineering-derived bias; as well as mitigation throughout the AI lifecycle. ISO/IEC 23282 Evaluation methods for accurate natural language processing (NLP) systems (<i>in development</i>) defines methods for evaluating NLP system accuracy around factual correctness, semantics, translation, question-answering performance, language understanding tasks and generative AI output quality. ISO/IEC TS 12791:2024 Treatment of unwanted bias in classification and regression machine learning tasks Methods for identifying, measuring and treating unwanted bias in ML systems developed to help organizations operationalise bias mitigation

AI failure category	Standards that can help
Hallucinations and fabricated outputs	ISO/IEC TS 42119-8 – Testing of AI Part 8: Quality assessment of prompt-based text-to-text systems that utilize generative AI (in development) focuses on assessing the quality of prompt-based generative AI systems and likely to become a key standard for evaluating hallucination rates and response quality. ISO/IEC 23282 Evaluation methods for accurate natural language processing systems (See above).
Customer-facing chatbot failure	EN 18286 AI Quality Management System includes post-market monitoring and clear requirements that help with monitoring, incident identification and recording, corrective action, learning from incidents, communication with affected stakeholders and requirements for understanding needs of interested parties.
Agentic AI failure	AIUC certification is the only agentic AI framework currently available. As a fast-developing area, international standards are in discussion around multi-agent interoperability, agent capability descriptions, agent orchestration frameworks, accountability and traceability, and multi-agent security and governance.

Interested in contributing toward standards development in AI to help shape AI safety and innovation?

Find how to get involved [here](#).

bsi

Get in touch
online

Explore more AI insights [here](#)

