Application Paper on the supervision of artificial intelligence

2 July 2025

About the IAIS

The International Association of Insurance Supervisors (IAIS) is a voluntary membership organisation of insurance supervisors and regulators from more than 200 jurisdictions. The mission of the IAIS is to promote effective and globally consistent supervision of the insurance industry in order to develop and maintain fair, safe and stable insurance markets for the benefit and protection of policyholders and to contribute to global financial stability.

Established in 1994, the IAIS is the international standard-setting body responsible for developing principles, standards and other supporting material for the supervision of the insurance sector and assisting in their implementation. The IAIS also provides a forum for Members to share their experiences and understanding of insurance supervision and insurance markets.

The IAIS coordinates its work with other international financial policymakers and associations of supervisors or regulators, and assists in shaping financial systems globally. In particular, the IAIS is a member of the Financial Stability Board (FSB), member of the Standards Advisory Council of the International Accounting Standards Board (IASB), and partner in the Access to Insurance Initiative (A2ii). In recognition of its collective expertise, the IAIS also is routinely called upon by the G20 leaders and other international standard setting bodies for input on insurance issues as well as on issues related to the regulation and supervision of the global financial sector.

For more information, please visit www.iais.org and follow us on LinkedIn: IAIS – International Association of Insurance Supervisors.

Application Papers

Application Papers provide supporting material related to specific supervisory material (ICPs and/or ComFrame). Application Papers could be provided in circumstances where the practical application of principles and standards may vary or where their interpretation and implementation may pose challenges. Application Papers do not include new requirements, but provide further advice, illustrations, recommendations or examples of good practice to supervisors on how supervisory material may be implemented. The proportionality principle applies also to the content of Application Papers.

International Association of Insurance Supervisors
c/o Bank for International Settlements
CH-4002 Basel
Switzerland
Tel: +41 61 280 8090

This document was prepared by the FinTech Forum in consultation with IAIS members.

This document is available on the IAIS website (www.iais.org).

IAIS

Contents

Acronyms

AI	Artificial intelligence
APIs	Application programming interfaces
DL	Deep learning
EU	European Union
EIOPA	European Insurance and Occupational Pensions Authority
G20	Group of Twenty
GenAI	Generative AI
HKIA	Hong Kong Insurance Authority
IAIS	International Association of Insurance Supervisors
ICP	Insurance Core Principle
LIME	Local Interpretable Model-Agnostic Explanations
LLM	Large language model
MAS	Monetary Authority of Singapore
ML	Machine learning
MRM	Model risk management
NAIC	US National Association of Insurance Commissioners
NYS DFS	New York State Department of Financial Services
OECD	Organisation for Economic Cooperation and Development
RAG	Retrieval augmented generation
SHAP	SHapley Additive exPlanations
SupTech	Supervisory technology

Executive Summary

1. The adoption of artificial intelligence (AI) systems is accelerating globally. For insurers, these developments offer substantial commercial benefits across the insurance value chain, for example by enhancing policyholder retention through personalised engagement, achieving significant cost reductions via increased efficiency in policy administration and claims management, or applying AI capabilities to improve risk selection and pricing.
2. However, with these advancements come risks that could negatively impact the financial soundness of insurers and consumers. For consumers, AI systems can, without safeguards, reinforce historic societal discrimination, increase concerns around data privacy and impede access to insurance. For insurers, the opaque and complex nature of some AI systems can lead to accountability issues, where it becomes difficult to trace decisions or actions back to human operators, and uncertainty of outcomes (particularly in a changing external environment). Addressing such concerns is paramount to maintaining trust and fairness in the industry.
3. Previous work by the International Association of Insurance Supervisors (IAIS) has affirmed that the current Insurance Core Principles (ICPs)[1] continue to be appropriate and relevant in managing these risks,[2] and at this stage no new standards are proposed. The objective of this Application Paper, therefore, is to support supervisors when applying the existing ICPs to promote appropriate and globally consistent oversight of the use of AI within the insurance sector.
4. This paper addresses a range of considerations related to the use of AI systems in insurance, with a focus on both consumer protection and prudential soundness. Its objective is to support supervisors, insurers and intermediaries by providing guidance on sound practices, and should not be interpreted as a prescriptive or exhaustive checklist. The overarching aim is that the use of AI does not adversely impact fair customer outcomes or undermine prudential standards. The practices outlined in this paper could be integrated into existing governance, risk management and control frameworks, avoiding the creation of new structures unless needed.
5. This Application Paper reinforces the importance of the ICPs, outlining how existing expectations around governance and conduct remain essential considerations for supervisors and insurers using AI. Furthermore, noting that AI can amplify existing risks, this paper emphasises the importance of continued Board and Senior Management education in order to establish robust risk and governance frameworks to support good consumer outcomes. Additionally, this paper notes that increasing application of AI can heighten the role of, and the risk from, third parties, like AI model vendors. Consistent with the ICPs, this paper reaffirms that insurers remain responsible for understanding and managing these systems and their outcomes.
6. Application Papers do not establish new standards or expectations but instead provide additional guidance to assist implementation and provide examples of good practice. This paper focuses on those requirements within the ICPs where systems could change the nature of the risk beyond those inherent in existing non-AI systems. The scope of the issues covered in the paper is

[1] The ICPs apply to insurance supervision in all jurisdictions regardless of the level of development or sophistication of insurance markets, and the type of insurance products or services being supervised.

[2] See IAIS, *Regulation and supervision of artificial intelligence and machine learning (AI/ML) in insurance: a thematic review*, December 2023.

deliberately limited to focus on those issues with higher risks. Furthermore, by placing emphasis on a risk-based and proportional approach, this Application Paper acknowledges the need to balance promoting innovation with minimising risk.

7. This paper leverages the work of other international organisations such as the Organisation for Economic Cooperation and Development (OECD) and the Group of Twenty (G20) to ensure a consistent approach to AI at the international level while considering sectoral specificities. Given the expected fast adoption of AI in the insurance sector, the IAIS will continue to monitor developments and will update material as appropriate.

1 Introduction

1.1 Context and objective

8. AI is a machine-based system that represents a series of techniques that aim to reproduce human intelligence by mimicking human cognitive functions such as perceiving, learning, exercising creativity and problem solving. There are different types of AI systems, with the common term "machine learning" (ML)[3] considered to be a subset of AI. Simpler AI systems focused on a specific task and using a fixed set of parameters applied to simple models are transparent and easy to understand, but lack flexibility. By contrast, AI systems, such as neural networks that are designed to imitate the functions and layered structuring of a human brain or deep learning (DL), are more complex and opaque, making it more difficult to interpret how a certain output was produced. A Generative AI (GenAI) system, such as a large language model (LLM), is an example of an AI system that combines the learning from two or more neural networks to understand and generate human-like text, graphics, sounds and videos, making these systems highly versatile for various tasks; however, it is difficult to trace why a certain output was produced.
9. The insurance industry has been using AI for some time within data analysis and predictive modelling. However, insurers are now actively testing and deploying AI systems more broadly throughout the insurance value chain, including policy administration and claims management, tailored customer engagement, enhanced risk assessment and fraud detection. Recent advancements in AI technology, specifically in GenAI, have unlocked a wide range of new potential applications across the insurance value chain.
10. AI systems bring numerous benefits for both insurers and policyholders, such as improved risk assessment and management, new products and services or cost reduction. Despite their benefits, these technologies can introduce new risks or increase existing ones, such as algorithmic bias or accountability issues linked to the opaque and complex nature of some AI systems. Box 1 below provides a more detailed overview of some of the potential benefits and risks related to AI systems.

[3] ML is an application of AI. It is the process of using mathematical models of data to help a computer learn without direct instruction. This enables a computer system to continue learning and improving on its own, based on experience. See Microsoft Azure, Artificial intelligence (AI) vs. machine learning (ML).

11. As these technologies become embedded in the sector's operations and decision-making, the need for effective oversight to ensure their ethical, fair, trustworthy and safe use is increasingly important. At the same time, it is also important that consumers and insurers can reap the benefits arising from AI systems. This Application Paper aims to find a balanced approach to the risks and benefits arising from AI systems, including by highlighting the need to take into account risk-based and proportionality considerations.

Box 1: Potential benefits and risks related to AI

Benefits

AI presents numerous potential benefits for both insurers and consumers. For example:

1. *Enhanced accuracy and granularity of risk assessments*: AI systems can enable insurers to analyse larger data sets from traditional and new data sources, from structured and unstructured formats, allowing them to develop more accurate and granular risks assessments. This allows insurers to take more informed decisions. For example, AI systems can be used to process satellite imagery to better underwrite natural catastrophe risks.
2. *Greater financial inclusion*: The development of more accurate and granular risk assessment can on the one hand exclude some consumers from insurance products and services (expanded on below), but on the other hand it can also facilitate the financial inclusion of some consumers. For example, the analysis by AI systems of data sets from telematics and wearable devices in motor and health insurance could result in young drivers and diabetes patients having access to more affordable insurance coverage.
3. *New tailored products and services*: AI systems can allow insurers to better understand the characteristics and needs of consumers, allowing them to develop more tailored products and services. For example, AI systems can be used to process data from wearable devices in health and motor insurance and develop personalised driving and lifestyle recommendations.
4. *Increased efficiency and lower costs*: Through the automation of certain tasks, AI systems can enhance the efficiency of certain processes that could eventually result in lower costs. For example, AI systems can be used to automate certain administrative tasks in the claims management area of the value chain such as the verification of invoices or doctor notes, freeing up staff to focus on higher-value tasks that require human expertise and judgment.
5. *Faster and more convenient processes*: Linked to the previous point, AI systems can also be used to develop more automated and faster processes. For example, by quickly analysing invoices and images, AI systems can speed up claims processing times. Furthermore, AI-powered chatbots available on a 24-hour basis from any location can conveniently support consumers during their first notification of loss.

Risks

Despite its benefits, AI can introduce new or enhance existing market conduct and prudential risks. The structure and guidance in this Application Paper is designed to support supervisors and insurers in managing the following potential risks:

1. *Data protection and security*: AI systems rely on the processing of large volumes of personal and non-personal data, increasingly sourced from secondary sources and not just provided by the customer. For example, LLMs using a retrieval augmented generation (RAG) process

vast amounts of potentially sensitive data outside of the core training data sources, which can raise concerns in respect of both data protection and confidentiality. Furthermore, some AI systems can potentially unmask anonymised data through inferences, ie deducing identities from behavioural patterns. Ensuring the privacy and security of customer information is crucial. Mishandling data can lead to breaches and legal consequences.

2. *Biased outcomes*: AI systems rely on identifying complex dependencies/correlations in the training data. Any biases in the training data or flaws in the system design will be inherited by the AI systems and can lead to reflecting and perpetuating socially biased outcomes. This can be particularly problematic for under-represented minorities that may historically have had limited opportunities in obtaining insurance (eg through historical prejudicial perception of higher risk). Biased outcomes increase the risk of poor policyholder outcomes, which in turn increase reputational (eg loss of business) and financial (eg regulatory fines) risks.
3. *Model risk/explainability*: Some AI systems are highly complex. Such complexity can reduce understanding and increase the uncertainty of model outcomes. For example, low explainability as to how decisions were derived can increase the risk of unwarranted or unlawful trends going undetected. Moreover, there are possibilities that models may not be able to respond to changes in the data. For example, if an AI system used in pricing and underwriting fails to adapt to a changing market, insurers may end up under or overcharging consumers, with potential consequences to their profitability and balance sheet.
4. *Uninsurability*: AI algorithms have the potential to assess risks in a very granular manner, which could potentially reduce risk pooling in insurance (reducing cross-subsidisation between policyholders), leaving certain riskier segments of society unable to access insurance at an affordable premium.
5. *Personalising profit margins*: Some AI systems can also be used to exploit the cognitive biases of consumers (including vulnerable ones), for instance by allowing insurers to extract additional revenue/profit based on consumer behaviours such as willingness to pay rather than risk. Different supervisors have observed that profit margins are often (much) higher for loyal customers than for newer ones. This typically means that loyal customers pay a higher premium without actuarial justification.
6. *Intellectual property infringement*: Certain AI systems learn from and rely on large quantities of external data sets that may inadvertently infringe existing patents or copyrights if there are no appropriate controls in place, which may lead to financial risks such as increased liability and litigation risks.
7. *Cyber security:* AI systems are vulnerable to data manipulation, data breaches and cyber attacks, which could prompt these models to make the wrong decisions. This includes risks from data poisoning, input attacks and model extraction.
8. *Concentration risks*: Insurers frequently purchase and/or outsource AI systems from a limited number of service providers. Failure in one of these service providers or a cyber attack affecting AI systems or the data sets provided by these service providers can affect the operational and cyber resilience of insurers and has the potential for systemic risk. For instance, if a number of insurers are using the same third party, significant IT outages affecting that provider can affect a large number of insurers.

See Section 1.3.1 for matters that are relevant for AI supervision but outside the scope of this Application Paper.

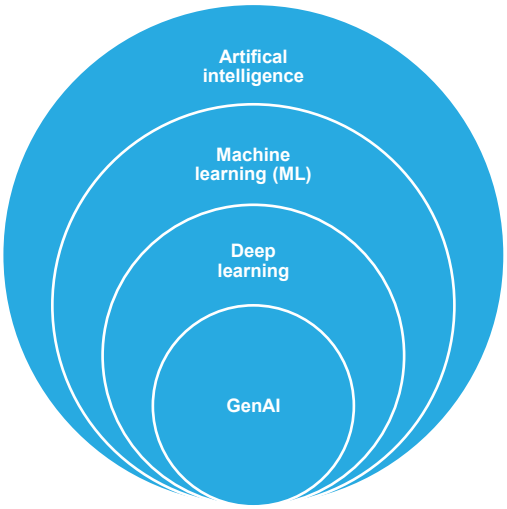
1.2 AI system definition

12. A clearly stated definition for AI offers several supervisory benefits, such as providing clarity and consistency of scope, helping to define the specific risks and assigning responsibility. There is no universal AI definition and, as the OECD highlights, there is no clear red line distinguishing between AI and non-AI machine-based systems (ie systems that do not use AI but may display some of the features of an AI system).
13. For the purpose of this Application Paper, and when considering the implications of AI, the following OECD definition[4] from 2024 provides a useful reference:

> An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.

14. This Application Paper, consistent with the definition above, adopts the reference to AI systems rather than simply AI, noting the OECD's observation that this is a more tangible and actionable concept; an AI system should be seen as a group of interacting or interrelated elements (eg the algorithm, data, assumptions etc) that form a unified whole. Importantly, autonomy and adaptiveness are two key elements in this definition that distinguish AI systems from more traditional mathematical models. The OECD definition is also sufficiently adaptable to capture fast-evolving areas such as LLMs that provide the underlying capabilities to enable GenAI[5] solutions.

Figure 1: AI and its different elements



[4] OECD, OECD AI Principles overview.

[5] A key aspect of GenAI is the ability to create new data and content. It can also respond naturally to human conversation and serve as a tool for customer service and personalisation of customer workflows. See Amazon Web Services, What is Generative AI?

15. The OECD definition provides a framework for distinguishing between AI and non-AI systems; however, it is important to note that current AI systems are, at their core, still mathematical models. Consequently, this Application Paper does not alter or supersede any existing requirements for monitoring and managing model risks, regardless of whether they are classified as AI systems. Instead, this paper provides guidance on novel or enhanced risks that arise from AI systems, which require particular attention when considering the implications for insurers and policyholders from prudential and conduct perspectives.
16. Furthermore, noting the blurred lines between AI and non-AI systems, insurers should consider whether any of the novel risks highlighted in this Application Paper are also present in any other model even if not defined as an AI system. Insurers should also consider whether setting out their own definition (whether aligned to the OECD or other variations) could clarify those models that may need more attention, as outlined in this paper. This would also provide clarity on existing models that are already in use and can continue to operate without the need for additional governance and risk management measures.

1.3 Scope and structure

17. The structure of this Application Paper is set out in Table 1 below and is designed to address the areas of governance and risk management identified as requiring particular attention when deploying AI systems. The areas cover both technical aspects, such as data governance and model validation, and those activities supporting an outcomes-based assessment. In aggregate, these are designed to address the risks highlighted in Box 1 above.

Table 1: Structure of the Application Paper – AI governance framework

Risk-based supervision and proportionality			
Governance and accountability	**Robustness, safety and security**	**Transparency and explainability**	**Fairness, ethics and redress**
• Risk management system • Corporate culture • Human oversight and allocation of management responsibilites • Use of third-party AI systems and data • Traceability and record keeping	• AI system robustness • AI system safety and security	• Explaining AI system outcomes • Explanations adapted to the recipient stakeholders	• Data management in the context of fairness • Inferred causal relationships in an AI systeem • Monitoring outcomes of AI systems • Adequate redress mechanisms for claims and complaints • Societal impacts of granular pricing

18. An ethical and responsible AI framework is achieved by a combination of governance and risk management measures set within the context of a broader business culture of responsible innovation and fair treatment of customers. Insurers need to develop a combination of governance and risk management measures that are appropriate for their specific AI use case.

For example, in certain circumstances the lack of explainability of a specific AI use case may be compensated by other measures such as increased human oversight and/or enhanced data management.

19. The purpose of this Application Paper is not to repeat traditional governance model risk management (MRM) requirements, but to focus on areas where AI systems could accentuate risks or where further guidance is seen as beneficial in addressing the unique characteristics presented by the deployment of an AI system. The ICPs covered by this Application Paper relate to: (i) managing model implementation and its ongoing use (ICP 8 (Risk management and internal controls) and ICP 16 (Enterprise risk management for solvency purposes)); (ii) appropriate oversight of and accountability for the model (ICP 7 (Corporate governance)); and (iii) managing model outcomes (ICP 19 (Conduct of business)).
20. The Application Paper supports supervisors in considering how the IAIS' ICPs should apply to both insurers and intermediaries insofar as an AI system is used in the various segments of the insurance value chain. The references to insurers in the paper should therefore be understood as applying to both insurers and intermediaries, unless explicitly stated otherwise.

1.3.1 Outside the Application Paper's scope

21. The scope of the Application Paper is deliberately limited; it is focused on managing risks related to the implementation and use of AI systems by insurers. As such, the following areas are out of scope:
    - Insurance-related risks associated with AI risks materialising within insured businesses, whether or not they are implicitly or explicitly covered by insurance products, such as risks arising from the use of AI in autonomous cars in the context of motor insurance, or risk arising from the use of GenAI to create fake claims; and
    - Investment-related risks resulting from the potential for financial markets to become more volatile due to AI-related risks.
22. The following aspects are also out of scope of this Application Paper, as they are not unique to AI and hence are covered in other guidance[6]:
    - Operational risks arising from other technologies such as those related to the implementation of cloud computing (note that such developments are an enabler; they are not unique to the implementation of AI); and[7]
    - Environmental issues arising from the high-energy consumption of AI systems, which may lead to an increase in greenhouse gas emissions and the consumption of natural resources.

[6] The IAIS is analysing AI use cases through its Global Monitoring Exercise, and more information will be in the 2025 GIMAR, expected to be published by end-2025.

[7] The IAIS has an extensive work programme on operational resilience. It finalised an Issues Paper on insurance sector operational resilience in 2023 and consulted on an Application Paper on Objectives for Operational Resilience in late 2024. In 2025, the IAIS will consult on a toolkit for operational resilience that will complement the objectives. These papers address (in part) issues and supervisory practices with respect to the provision of third-party IT services, including the use of the cloud.

Table 2: Overview of ICP standards covered

ICP	Topic	ICP	Topic
1.4	Objectives, powers and responsibilities of the supervisor	8.6	Actuarial function
2.10	Supervisory resources	8.7	Internal audit function
5.2	Competence and integrity	8.8	Outsourcing oversight
7.1	Appropriate allocation of oversight and management responsibilities	16	Enterprise risk management for solvency purposes
7.2	Corporate culture	19.0	Fair treatment of customers
7.3	Delegation of responsibilities	19.2	Policies, processes and business culture of fair treatment of customers
7.4	Board member responsibilities	19.7	Information for consumers
7.5	Duties related to risk management and internal controls	19.10	Claims handling
8.1	Systems for risk management and internal controls	19.11	Complaints handling
8.4	Risk management function	19.12	Protection and use of customer information
8.5	Compliance function		

1.4 The role of supervisors and supervisory tools

23. ICP 1 (Objectives, powers and responsibilities of the supervisor), notably ICP 1.4.1, states that it is "important that supervisory responsibilities, objectives and powers are aligned with actual challenges posed by the insurance market to effectively protect policyholders, maintain a fair, safe and stable insurance market and contribute to financial stability".

24. ICP 2 (Supervisor), notably ICP 2.10, states that the supervisor has "sufficient resources, including human, technological and financial resources, to enable it to conduct effective supervision". This includes providing adequate training for staff to ensure their knowledge, skills and supervisory practice remain up to date.

25. Considering AI systems' developments and their broad deployment, supervisors play an important oversight role and will need to understand these developments in order to undertake effective supervision. Specifically, supervisors should consider how they intend to identify, assess and monitor the challenges that arise from the increasing deployment of AI systems, while developing and maintaining their technical supervisory capabilities in this area. Supervisors may wish to consider the following tools and approaches to assist them:

    - *Develop training/knowledge*: Over time, supervisors should foster a deep understanding of AI technologies to effectively oversee their use and challenge their outputs when the need arises. This can be achieved by taking a forward-looking approach to supervisory resources and their training needs. Authorities should provide training for supervisors, covering, for example, answers to (i) what is an AI system; (ii) how is it deployed; and (iii) what are the potential risks. Any such training should be regularly reviewed and updated given the developing nature of AI systems. As AI use increases, so too should the available training for supervisors. Depending on the pace of AI development, authorities should consider setting up centres of expertise that serve as hubs for AI research (including collaboration with

industry experts and academic institutions), knowledge sharing, monitoring the industry's progress, creating case studies and embedding lessons learnt.

- *Cooperation and coordination with insurers and other authorities (both at the jurisdiction level and internationally)*: Depending on the existing supervisory architecture in a jurisdiction, there may be more than one authority involved in the supervision of the use of AI systems by insurers. This may include conduct authorities, prudential authorities, data protection authorities or other relevant agencies. Existing cooperation channels, forums or committees could be used or enhanced, or new ones established, to encourage the sharing of experiences and knowledge. Since AI trends are likely to be global in nature, there are significant benefits for supervisors engaging at an international level. They can share supervisory experiences and knowledge, ensuring the transfer of information on techniques, methods and supervisory approaches. At the international level, the IAIS, via the FinTech Forum, provides a mechanism for information exchange amongst supervisors, and the IAIS works closely with the other standard-setting bodies on these issues.
- *Use of innovation facilitators*: Sandboxes and innovation hubs can support a test environment allowing supervisors to explore different approaches to supervision and can help support the development of rules or conditions supervisors may want to put in place. Sandboxes also help to promote dialogue and communication, and enable supervisors to communicate supervisory expectations.
- *Use of surveys*: Targeted supervisory surveys can help (i) identify the variety of differing AI system use cases; (ii) inform a risk-based approach to supervision; (iii) provide transparency to the market on areas of interest; and (iv) identify AI concentration risks.
- *Use of supervisory question banks*: Developing a comprehensive supervisory question bank[8] can support consistency in review and decision-making. Such a question bank could also be used to support resource planning, ensuring that the appropriate mix and quantum of technical and conduct-related expertise is available to support any review. [9]
- *Learning from supervisory technology (known as SupTech)*: Many authorities are developing and deploying new AI tools designed to support effective supervision, such as outlier detection using AI to identify insurers with potential for elevated prudential risk. Supervisory knowledge of SupTech tools using AI can be enhanced through dialogue with IT departments, facilitating understanding of the issues and complexities identified when deploying such tools.

[8] Question banks are sets of questions used by supervisors for engagement with insurers on specific topics. They provide supervisory teams with a consistent way of engaging with insurers and help supervisors to understand the level of knowledge across the sector on a particular issue.

[9] In 2020, the NAIC published a Regulatory Review of Predictive Models White Paper, which includes appendices on specific ML model types. One of the appendices was on tree-based models such as gradient boosting machines.

2 Risk-based supervision and proportionality

26. This Application Paper should be understood within the context of risk-based supervision and the proportionality principle, as described in the Introduction to the ICPs. Specifically, supervisors can adjust their implementation of the ICPs to achieve the outcomes stipulated in the Principle Statements and Standards.
27. When considering the advice, illustrations, recommendations and examples of supervisory good practices presented in this paper, it is important to bear in mind that the principle of proportionality underpins all ICPs. This principle also informs the governance and risk management measures outlined in the Application Paper.

2.1 Risk-based supervisory approach

28. Risk-based supervision implies that supervisory activities and resources are allocated to insurers, lines of business or market practices in line with the level of risk to policyholders, the insurance sector or the financial system as a whole. In the context of AI systems, it is acknowledged that there are different types of AI systems and use cases carrying different levels of risks. For example, an AI system used for efficient document retrieval will carry less risk than one determining the claim payouts to policyholders. And supervisory expectations will be higher for the use of AI systems affecting retail customers compared to commercial customers as is consistent with the ICPs[10].
29. A framework that distinguishes between various levels of risk can support both the application of proportionality as well as risk-based supervision. It can ensure that supervisory resources are allocated to higher-risk AI use cases that present the greatest potential for market conduct and/or prudential risk.
30. Table 3 provides illustrative guidance on certain criteria or characteristics that supervisors and insurers could consider when assessing and assigning a level of risk to an AI system. Each criterion is grouped into two broad categories that capture whether the assessment requires an evaluation of outcomes or a technical evaluation of the underlying model. This recognises that the supervision of AI systems requires a combination of outcomes (eg assessing implications for policyholders, especially retail customers) and technical-focused activities (eg assessing data/model validation).
31. This is not a checklist; rather, it is intended to support the development of suitable risk-based frameworks that can reflect the specificities of the legal, societal and jurisdictional aspects in which insurers operate. Note that the large number of listed criteria reflects the variety of ways AI systems are or could be deployed within organisations. The relevance and importance of each criterion should reflect the AI systems that are being considered and deployed across an insurer's business model.

[10] General expectations on fair treatment of customers are set out in ICP 19.1 and in particular ICP 19.2.2, which notes that "Proper policies and processes dealing with the fair treatment of customers are likely to be particularly important with respect to retail customers, because of the greater asymmetry of information that tends to exist between the insurer or intermediary and the individual retail customer".

Table 3 – Examples of criteria to assess the risks of AI systems

	Focus	Criteria/ characteristics	Explanation
Outcomes-related	Policyholders	Application	The nature of the decision being made, including the potential implications for existing or prospective policyholders. For example, AI systems within underwriting and claims processes could be deemed higher risk, reflecting a direct link to consumer outcomes.
		Unlawful discrimination	The extent to which the AI system has had or has the potential to unlawfully discriminate against protected classes.
		Fairness	The extent to which the AI system is engaged in responsible stewardship in pursuit of equitable treatment of consumers.
		Volume/type of customers affected	The extent to which the AI system has had or has the potential to have an adverse impact on a plurality of customers, in particular vulnerable ones.
		Line of business	The extent to which the AI system is used in a line of business that is important for the financial inclusion of customers and/or when there is a legal requirement to obtain such insurance.
		Reversibility and redress	The extent to which consumers have a way to inquire about how the AI system arrived at its decision/outcome and the extent to which an outcome is easily reversible, with an effective process for redress where appropriate.
	Insurers	Critical insurance function	The extent to which an AI system can cause a disruption to core business activities, eg issuing policies or managing claims.
		Financial impact	The extent to which a failure of an AI system could result in a material impact on the financial commitments of an insurer.
		Legal impact	The extent to which the failure of an AI system could result in a violation of legal commitments with the potential for critical impact on an insurer.

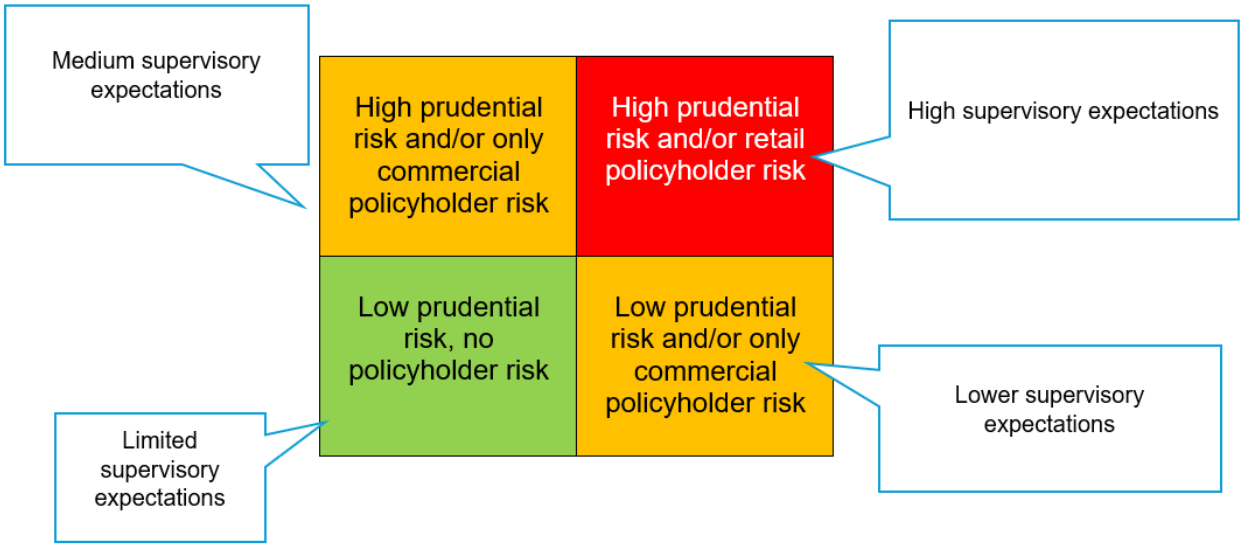
Model-related	Architecture	Knowledge & resources	The extent to which the insurer has the necessary knowledge and resources in place for the selected AI system to comply with all applicable insurance standards, laws and regulations, including privacy and data security concerns.
		Adaptability	The extent to which the AI system has the ability to recalibrate itself, thereby changing the underlying model structure, as new information becomes available. Such adaptability could be considered to increase the risk of unintended biases as the model deviates from the latest signed-off model version. There is also an important time dimension, with additional consideration needed for models that update and adapt in or near "real time".
		Data quality	The extent to which appropriate training data that is sourced and utilised for model-building is accurate, complete, unbiased, and representative for the nature of the population affected by the implementation of AI systems.
		Transparency/ explainability	The extent to which the AI system's outcomes/decisions can be understood, explained and documented in a meaningful way, revealing the nature of the input data being used, the purpose of the data and the potential consequences of risk to relevant stakeholders (eg consumers, Senior Management, auditors, supervisors etc), for the purpose of improving the public's confidence in AI while protecting the confidentiality of proprietary algorithms.
	Implementation	Autonomy	The extent to which humans are involved in the final decision-making process.
		Missing information	The nature and impact of missing data, such as missing completely at random, systematically missing due to factors correlated with risk, missing due to limitations of incomplete data from third parties or missing due to insured's or insurer control.
		Third-party reliance - model/system - secondary data	The extent to which an insurer's business or management decisions are reliant on and influenced by a concentrated number of AI system service providers or other third-party data providers supporting the deployment of AI systems.

2.2 Application of the principle of proportionality

32. In line with the principle of proportionality, supervisors should require that insurers implement governance and risk management measures that are commensurate with the risk profile of the specific AI system in use. Higher-risk AI applications should be subject to more robust oversight and controls, whereas lower-risk systems may warrant a more proportionate approach. Risk assessments should therefore inform the depth and breadth of measures adopted, so supervisory expectations remain both risk-sensitive and outcomes-focused.

33. The diagram below illustrates how an AI system risk assessment (potentially leveraging criteria set out in Table 2) could be integrated into a governance framework that supports proportionate supervision. Indeed, such a risk assessment framework allows insurers and supervisors to identify which AI use cases pose higher risks and accordingly develop more rigorous and stringent governance and risk management measures for those AI systems that pose the greatest risks.

Figure 2: Risk-based approach and proportionality



34. It is important to recognise that a responsible governance framework is achieved by a combination of governance and risk management measures, rather than relying on a single measure alone. For example, certain AI systems used to process images, text or videos may inherently have low levels of explainability, but in view of their benefits the low explainability can be compensated with alternative governance measures such as human oversight or data management. In the application of the principle of proportionality, insurers and supervisors should adopt a holistic approach to AI governance by assessing the combination of governance and risk management measures jointly.

3 Governance and accountability

3.1 Introduction

35. The development of AI systems involves numerous inherent features that are particularly relevant to governance and accountability. Most notably:
    - *Rapid technological advancements*: The newness and swift pace of change in AI technologies, coupled with their diverse application in insurance-related contexts, present unique and evolving challenges for risk management.
    - *Lack of AI expertise*: In this emerging field, there is often a shortage of skills, knowledge and expertise, including at the Board level, in both the development and proper usage of AI systems.
    - *Strong business incentives*: In many areas, AI-driven innovations are perceived as critical to maintaining an insurer's competitive position and unlocking further business success. Risk management and governance measures need to evolve at a similar pace to ensure long-term success.
    - *Potential for broader societal implications*: AI systems can make rapid evaluations based on detailed, granular information, often down to the individual consumer. This capability highlights the need for corporate strategies to balance profit maximisation with good consumer outcomes.
    - *Potential lack of centralised accountability:* The broad implementation of AI systems across various functions in the insurance process chain can lead to dispersed accountability for managing risk.
36. There are a number of ICPs that cover topics relevant to governance and accountability. These are:
    - *ICP 8 (Risk management and internal controls)*: "The supervisor requires an insurer to have, as part of its overall corporate governance framework, effective systems of risk management and internal controls, including effective functions for risk management, compliance, actuarial matters and internal audit.";
    - *ICP 16 (Enterprise risk management for solvency purposes)*: "The supervisor requires the insurer to establish within its risk management system an enterprise risk management (ERM) framework for solvency purposes to identify, measure, report and manage the insurer's risks in an ongoing and integrated manner.";
    - *ICP 7 (Corporate governance)*: "The supervisor requires insurers to establish and implement a corporate governance framework which provides for sound and prudent management and oversight of the insurer's business and adequately recognises and protects the interests of policyholders."; and
    - *ICP 5 (Suitability of persons)*: "The supervisor requires Board Members, Senior Management, Key Persons in Control Functions and Significant Owners of an insurer to be and remain suitable to fulfil their respective roles."
37. This section covers the additional areas within these ICPs that, due to the inherent characteristics of AI systems, require specific attention.

3.2 Risk management systems

38. ICP 8.1 states that "The supervisor requires the insurer to establish, and operate within, an effective and documented risk management system, which includes, at least: a risk management strategy that defines the insurer's risk appetite; a risk management policy outlining how all material risks are managed within the risk appetite; and the ability to respond to changes in the insurer's risk profile in a timely manner."
39. The management of material AI related-risks can be set out in existing risk management policies, such as within an existing model risk management policy or an AI-specific policy. Either way, a clear articulation and common understanding across control functions (including risk management, compliance and internal audit) of what constitutes AI-related risk, and the development of risk assessment criteria are important. Section 2 provides possible risk characteristics that, together with consideration of potential adverse outcomes (set out in Table 3), could support insurers in developing a risk framework and risk appetite statement, as well as metrics to support the monitoring of AI-related risks that could be regularly reviewed.
40. When adopting AI systems, the main requirements for each of the control functions (as set out in ICPs 8.4–8.7) remain appropriate. Nevertheless, insurers and their supervisors should regularly assess whether the skills, resources and capabilities within these functions are aligned with the evolving advances in AI systems and the level of deployment.

3.3 Corporate culture

41. Under ICP 7.1, supervisors should require Boards to "ensure that the roles and responsibilities allocated to the Board, Senior Management and Key Persons in Control Functions are clearly defined so as to promote an appropriate separation of the oversight function from the management responsibilities; and provide oversight of the Senior Management." In adopting AI systems, insurers should ensure that activities are consistent with their corporate culture and that fair treatment of customers is an integral part of that culture, with policies and processes properly embedded to support this objective in line with ICP 19.2 ("The supervisor requires insurers and intermediaries to establish and implement policies and processes on the fair treatment of customers, as an integral part of their business culture").
42. When implementing a risk-based approach to AI risk management, the Board should promote a corporate culture for fair and ethical outcomes, in alignment with ICP 7.2. Examples of responsible AI use that insurers could adopt include (see also Section 6 below):
    - Defining its approach to fairness and overseeing the implementation of norms for responsible and ethical behaviour, specifically ensuring these norms are made clear to those employees that are involved in the purchase, development, validation, implementation and audit of AI systems, as well as those who use AI systems in their work. Regular monitoring and training on these norms should also be carried out. Tone from the top is important to establish these norms as part of the corporate culture.
    - Clear accountability for setting expectations with regard to AI systems so the output generated by these systems is fair, explainable, unbiased and ensures adequate policyholder protection.
    - Enabling strong compliance and risk functions and promoting a constructive feedback and remediation culture. This will mean that risk management approaches are robust, and designed and implemented in parallel to the adoption of new AI systems (not lagging behind), and any issues that may arise are identified and acted upon early.

3.4 Human oversight and allocation of management responsibilities

43. The development, implementation and oversight of AI systems throughout their entire life cycle should not alter supervisory expectations. For example, Boards should continue to ensure that insurers have a well-defined and documented governance structure that provides effective separation between oversight functions and management responsibilities.

44. Four ICPs are particularly relevant here:

- ICP 5.2 emphasises that "Board Members (individually and collectively), Senior Management and Key Persons in Control Functions possess competence and integrity" in their roles. Meanwhile, ICP 5.2.1 notes that "Competence is demonstrated generally through the level of an individual's professional or formal qualifications and knowledge, skills and pertinent experience within the insurance and financial industries or other businesses."
- ICP 7.3, because the Board, with its collective expertise, is tasked with challenging Senior Management's decisions in the context of AI, so a robust oversight of AI systems is required.
- ICP 7.4, which states that individual Board members are required to exercise due care, diligence, independent judgement, and objectivity in their decision-making processes.
- ICP 7.5, which notes that a Board's oversight extends to the design and implementation of risk management frameworks and internal controls. In this context this includes to address the insurer's use of AI systems. This collective accountability means that insurers operate responsibly, mitigate risks, and align with regulatory expectations.

3.4.1 Board and/or Senior Management

45. There are a number of inherent characteristics of AI systems that necessitate particular attention, which include but are not limited to:

- *Defining responsibility for the AI system throughout its life cycle (design, approval, development, procurement, deployment, monitoring and decommissioning)*: This could consider the use of a detailed responsibility matrix outlining roles at each stage and a structured handover process to maintain accountability. Specific areas for careful consideration include where a data scientist may be responsible for initial deployment, but where responsibility may shift to the business areas as the AI system updates and adapts to new policyholder information. Furthermore, in recognition of the pace of change with regard to AI systems, Senior Management should review policies and processes regularly to confirm alignment with relevant regulations, industry standards and best practices for responsible AI.
- *Establishing appropriate baseline expertise*: Where AI is used for important decision-making, Board members and/or Senior Management should have an understanding of its risks and limitations in order to effectively challenge its output and understand its impact on the business strategy. This should also include awareness of the threats and opportunities of AI within the insurance sector and the extent to which these could have implications for an insurer's business strategy and viability. For an insurer that makes heavy use of AI in processes that significantly affect consumer outcomes, it is essential to have sufficient Board expertise to consistently deliver effective AI solutions that safeguard against consumer harm. More broadly, Senior Management should be confident that effective training is cascading throughout the insurer so that all staff are aware of the risks of AI and understand their role in addressing these risks. Noting that AI is a fast-developing area, the Board and Senior Management should consider regular training to acquire, maintain and enhance their

knowledge and skills in order to provide objective and robust scrutiny of the deployment of AI systems.

- *Achieving effective human oversight*: This should include any prerequisite training for those tasked with providing human oversight; for instance, with respect to data sets, ensuring training in false, biased, unethical or unfair outcomes detection and ensuring that those individuals who provide oversight are independent from the model development process in order to maintain objectivity (ie a second line). In this regard, it is important that key people in control functions have the appropriate knowledge and skills to understand and recognise the potential business, human and societal implications. It is also important that the insurer's corporate culture allow for such issues to be raised and then acted upon.
- *Managing the limitations of human oversight*: Many AI systems are purchased from third-party service providers. Such systems are frequently characterised by limited access to the underlying infrastructure, code and source of the training data. This can challenge the effectiveness of human oversight. In addition to standard risk management strategies (such as due diligence and third-party assessments), insurers should examine, where applicable, the necessity of system redundancy, oversight of inputs and outputs to AI systems using mechanical controls and so-called kill switches that would cause the AI system to stop functioning under certain pre-specified conditions. Senior Management may establish performance indicators/metrics specific to AI systems that align with the insurer's risk appetite and are regularly reported to the Board.

3.4.2 Additional Senior Management responsibilities

46. Senior Management is responsible for the day-to-day management of the insurer, which includes its day-to-day operations, risk management, compliance and fair treatment of customers. In relation to AI systems, it is crucial that Senior Management establish clear procedures for addressing specific challenges that are more difficult to manage when deploying AI systems. These procedures should ensure effective governance, including mechanisms for monitoring AI performance, detecting biases and implementing corrective actions promptly. With respect to AI systems, this should include establishing procedures for addressing issues known to be harder to achieve when deploying an AI system. For example:
    - Achieving clear lines of accountability by considering who holds ultimate responsibility for the model;
    - Ensuring human oversight provides a robust and objective control;
    - Providing transparency on objectives, as well as how short- and long-term rewards are balanced where the AI system includes reinforcement learning;
    - Achieving effective communication strategies when the underlying system is by nature opaque and complex;
    - Establishing appropriate record keeping, particularly when the basis of future decisions could change autonomously; and
    - Setting clear guardrails on when an AI system can or cannot be deployed.

3.5 Use of third-party AI systems and data

3.5.1 Third-party oversight

47. Governance and management of third-party risk is likely to become increasingly important as new AI models such as GenAI are adopted by insurers, and they need to be embedded alongside existing governance measures.
48. Board and/or Senior Management collectively retain responsibility for appropriate oversight of third parties conducting activities for the insurer and any nth-parties that the third parties rely on, including as part of outsourcing arrangements. The insurer should assess whether acquiring, using or relying on AI systems developed by a third-party (or nth party) constitutes an outsourcing of critical services (as set out in ICP 8.8). The IAIS glossary defines outsourcing as "an arrangement between an insurer and a service provider, whether internal within a group or external, for the latter to perform a process, service or activity which would otherwise be performed by the insurer itself".
49. Where an insurer uses third parties or outsourcing and the providers use AI systems, the same level of oversight should be expected as if the insurer had developed the AI system (ICP 8.8). However, third-party service providers also have a role to play in the implementation and adoption of responsible and trustworthy AI systems. Accordingly, insurers should involve third parties (and nth parties), as relevant, in their assessment of potential limitations and risks of the use of third-party AI systems and data.
50. Taking into account the intellectual property rights of third parties, insurers should obtain adequate information and reassurances from third-party (and nth-party) service providers (for example, via clauses in the contracts between the insurer and its third-party service providers) about the characteristics, capabilities, appropriate fitness for purpose and limitations of AI systems they outsource where they are critical services. For example, insurers can consider including the specific provisions in contracts with third-party AI providers, such as, but not limited to: (i) transparency requirements regarding model architecture and training data sources; (ii) regular independent auditing rights; (iii) commitments to ethical AI principles aligned with the insurer's values; (iv) service-level agreements for ongoing performance monitoring; and (v) clearly defined processes for model updates and retraining.

3.5.2 Third-party concentration risks

51. The market for AI services may be concentrated, with potential implications for the market power of individual providers. Insurers should make regular assessments of the extent to which the insurer's reliance on an AI service provider may pose a risk to their business. They should consider the related operational risk and the steps that could be taken to mitigate this risk, including a comprehensive exit plan that should consider the potential circumstances and triggers under which such a plan may need to be enacted.[11] Concentration risks in third-party AI services could also cause concern for supervisors, given the potential for systemic implications if there was an operational issue with that third party. Supervisors should monitor such concentrations and be aware of the breadth of insurers that could be impacted. Moreover, insurers should develop and regularly test contingency plans that detail specific steps to maintain

[11] Objective 2.8 of the IAIS *Draft Application Paper on operational resilience objectives and toolkit* sets out considerations for insurer on the management of third and n-th party relationships.

business continuity in the event of a third-party AI service provider failure, including transitioning to alternative systems or manual processes where necessary.

3.6 Traceability and record keeping

52. For reproducibility and traceability of the AI system, supervisors should assess how insurers implement mechanisms that can track data sources used in training AI systems and the processes involved in content generation. Tools like data provenance frameworks and model cards for model reporting can be used to document and trace the life cycle of AI systems, including the data sets used, training processes and any modifications made to the models over time. Event logs can be used to record all meaningful activity associated with the AI system. These reports should be made available to supervisors and auditors to enable them to assess and challenge the decisions of AI systems. This practice would also support and facilitate access to adequate redress mechanisms (see Section 6.5 and the Annex below).
53. Given the principle of proportionality, for high-risk AI applications it is recommended to maintain repositories that contain all deployed models within the organisation. An example of the main attributes that could be recorded for each AI system (whether developed internally or outsourced) is provided in the Annex.

4 Robustness, safety and security

4.1 Introduction

54. In contrast to traditional systems that typically rely on explicit human-engineered rules and logic, AI systems, and especially foundation models, learn from very large data sets. They recognise patterns and generate outputs by analysing information across different domains. Unlike traditional models, AI systems can tackle complex tasks with intricate patterns and highly complex non-linear relationships. Furthermore, some AI applications can continuously update their understanding and predictions with new data and can adapt to changing circumstances. These differences highlight the need for additional safeguards around model validation (particularly where a model adapts over time) and the underlying data storage and use.
55. This section covers the technical aspects that insurers and supervisors should consider when assessing the risk management of an AI system and the underlying data. It covers the assessment of the robustness of the model, safeguarding of policyholder information and security of the AI system.
56. This section supports the implementation of ICP 8 and ICP 19. Significant amounts of information collected, held or processed by insurers represent customers' financial, medical and other personal information. Security over such information is extremely important; hence, safeguarding personal information on customers is one of the key responsibilities of the financial services industry.

4.2 AI system robustness

4.2.1 Performance

57. Insurers may wish to regularly assess, evaluate and document the performance of their AI systems. The performance measures could consider the underlying objective and the known

model and data limitations. The performance metrics (accuracy, recall, precision etc.) could depend on the nature of the data and the context and objectives of the AI system, for example the use of lower thresholds of acceptance for AI decision errors that directly affect policyholders. In addition to performance metrics, insurers may also wish to consider the following when testing for the robustness of an AI system, where applicable:

- *Out-of-sample testing to discover potential overfitting*: Test results on data with known outcomes with data not used during training.
- *Benchmarking*: Check against other models or expected results.
- *Sensitivity analysis*: Understand changes in outputs resulting from small changes in the inputs.
- *Adversarial testing*: Subject the AI system to invalid inputs to understand how the model behaves.
- *Stress testing:* Assess the system's performance under extreme conditions, such as high workloads, data spikes or sudden changes. Robust systems should handle stress without compromising accuracy or stability.
- *Data diversity testing and use of synthetic data:* Validate the AI system performance across diverse data sets. Where historical data may not be complete, insurers may cautiously consider the use of synthetic data, subject to robust validation and governance processes to maintain reliability.
- *Edge case testing:* Investigate rare or unusual cases that might not be adequately covered during regular testing. Creating and maintaining a repository of these edge cases can reveal vulnerabilities or unexpected behaviour and supports continued evaluation of the AI system.
- *Concept drift:* Monitor the AI system's performance over time. Concept drift occurs when the underlying data distribution changes. This includes identifying when a meaningful deviation in the system has occurred, which may affect model performance or fairness. Regularly retrain and validate the model to maintain robustness.
- *Interoperability testing:* Ensure integration with existing systems, application programming interfaces (APIs) and third-party services. This should consider the entire ecosystem that is influenced by AI system decisions. For example, rigorous API versioning control with backward compatibility is needed to maintain system stability during updates.

The level of rigour and frequency of these performance tests should reflect the potential impact of the AI system. For AI systems that pose higher risks, more in-depth testing and monitoring may be necessary.

58. Robustness testing could be an ongoing process, and insurers may wish to adapt their strategies as the technology evolves. Furthermore, implementing automated monitoring tools that trigger alerts when significant changes in data distribution are detected supports a proactive approach and timely model updates. Expected outcomes could be defined before seeing the results. Where an insurer uses a third-party AI system, any material findings from the robustness assessment could be shared with the provider or developer so that corrective measures can be addressed promptly, where applicable.

4.3 AI system safety and security

59. Deploying AI systems involves several safety and security concerns that need to be addressed to protect sensitive data and comply with regulations. Cyber security risks can originate from

inefficiencies in various phases throughout an AI system's life cycle, for instance in design (eg security architecture not adequately designed or insecure data storage and transmission), development (eg code vulnerabilities) and deployment (eg delayed security patches). Insurers should implement advanced security measures against potential threats, in particular against cyber attacks (see also the UK case study in the Annex). This could involve developing regular adversarial testing and continuous monitoring for anomalies to identify potential threats like data poisoning and model inversion attacks. Additionally, automated alerts potentially supported by AI solutions may also strengthen insurer's ability to detect significant deviations in AI behaviour, allowing for swift corrective actions.

60. Malicious actors can attempt to alter AI systems' use, output, performance or behaviour, or exploit system vulnerabilities by compromising model security. There are a number of tools, such as intrusion detection systems, threat intelligence platforms and endpoint detection and response solutions, that detect and respond to threats in real time, ensuring vulnerabilities are addressed swiftly. By capturing security measures with their AI systems, insurers can proactively defend against sophisticated attacks and maintain their systems and data. In addition, security assessments should account for AI-specific vulnerabilities, including data entry point attacks and prompt engineering exploits. Penetration testing, red teaming and other forms of stress-testing should be carried out by trained experts under appropriate confidentiality.
61. Maintaining up-to-date security practices is critical as threats evolve. Regular updates of security tools for AI systems, alongside continuous staff training on new risks, are essential.
62. Additionally, in common with other processes, insurers should put in place effective backup and recovery solutions to maintain business continuity for insurers, especially where AI systems provide critical functions. Where appropriate, automated alerts can be employed to detect significant anomalies or deviations. However, insurers should determine the suitability of such alerts based on the risk profile and scale of their AI systems.
63. ICP 8.8 sets out clear expectations about the need for insurers to maintain "at least the same degree of oversight of, and accountable for any outsourced material activity or function". Therefore, when using AI systems where third-party providers are involved, insurers remain responsible. They should carry out a security risk assessment to take appropriate steps to mitigate security risks, including assessment of how the data is transmitted, stored and encrypted.[12] Examples of AI-related security risks include malicious inputs aimed at triggering unintended outputs (prompt injection), data poisoning or adversarial attacks tailored to exploit AI model weaknesses.

4.3.1 Segmentation and compartmentalisation

64. As a mitigation against risks from cyber attacks, insurers may consider implementing a segmentation and compartmentalisation strategy within the AI system and its purpose-built models as an additional control measure (already common in cyber risk frameworks). At the same time, integrated or connected data sets can bring efficiencies, for example in data lineage or cleaning, so insurers should balance segmentation with operational needs. Isolating critical components would limit the impact of any single point of failure, thereby enhancing the system's resilience against potential attacks or data poisoning.

[12] The IAIS' Issues Paper on insurance sector operational resilience sets out more details on issues including cyber resilience and third-party outsourcing. Additionally, the US National Institute of Standards and Technology provides a standard for understanding cyber attacks in its publication Adversarial machine learning: A taxonomy and terminology of attacks and mitigations.

Box 2: Additional considerations for GenAI and LLMs

The use of GenAI and LLMs is expanding rapidly, enabling various applications in text, image, or code generation. Examples include improving the level of engagement of chatbots in providing advice and/or recommendations to consumers or sales agents, producing different regulatory filings, speeding up claims handling processes, enhancing fraud detection, and reducing the time spent by actuaries, underwriters and claims adjusters on administration.

Due to their specific nature and complexity, GenAI and LLMs could also bring a number of new risks or enhance existing ones, such as potentially providing incorrect or inaccurate advice to consumers or to sales agents (the so-called hallucinations), biased outputs as a result of the use of biased data sets on the internet, or lack of explainability. The fact that several service providers reportedly do not disclose the data sets they have used to train their models also makes it difficult for insurers to perform adequate data management processes (eg to remove biases).

Developing and implementing GenAI tools also involve several risks related to copyright and intellectual property rights that can lead to legal disputes and liability issues. For example, data scraping raises concerns about whether data creators should be compensated. Moreover, the ownership of outputs can also raise complex legal ambiguities, since in various jurisdictions there are provisions that provide a unique category for computer-generated works. Plagiarism and originality issues are another example, since LLMs can generate content that closely mimics existing works.

From a different perspective, GenAI tools can also produce fake reports or images (eg a picture of a car with false damage) that could be used to make fraudulent claims. GenAI tools and foundation models also increase the capabilities of hackers to carry out cyber attacks (see also the Singapore example in the Annex).

While the inherent complexity and characteristics of GenAI and LLMs make them unique amongst AI systems, the AI governance measures described in this Application Paper are equally applicable to them. Supervisors should ensure that insurers develop adequate governance measures to address their limitations in terms of explainability or data management, for instance by gathering sufficient reassurance from third-party service providers or by monitoring the outcomes of AI systems. Insurers should be mindful of the limitations of such tools, in particular with regard to so-called hallucinations, which can be mitigated by having a human validate the outcomes. Such governance measures could also include regular training and workshops for insurers on intellectual property rights and emerging legal trends, and the establishment of a dedicated task force to continuously monitor and address AI-related legal risks.

Supervisors should also require that insurers deploying GenAI tools and LLMs stay informed about these risks and manage their legal risk as they navigate the complex landscape of intellectual property rights in the context of AI-generated content.

5 Transparency and explainability

5.1 Introduction

65. Some AI systems are seen as “black boxes” due to their complex internal functioning; they can learn from data with various levels of autonomy, making it challenging to explain how decisions are reached (eg why a consumer’s insurance application has been rejected or accepted), or the

role/weights of specific variables (eg a consumer's address, age, driving experience etc) or combinations of variables in the outcome of the AI system. This is particularly the case when the AI system is trained with large data sets (also known as big data). Moreover, as with human decision-making processes, there may be inherent biases or a lack of transparency.

66. Transparency and explainability are key to building trust and ensuring accountability for understanding the unique risks to consumers, such as the potential for unlawful discrimination. This is relevant because ICP 19 states that supervisors require insurers and intermediaries "act with due skill, care and diligence when dealing with customers". ICP 19 also highlights the importance of treating customers fairly and providing clear, timely and adequate information allowing them to make informed decisions.[13]
67. A lack of understanding of the functioning of an AI system may also have implications from a prudential perspective. For example, if an AI system is used in underwriting and inadvertently fails to price risk segments accurately, the insurer could potentially acquire risks at a premium level that is insufficient to meet the future claims cost. To prevent this, it is important to have effective systems of risk management and internal controls in line with ICP 8. Given insurers often rely on AI systems developed by third parties, adequate oversight should extend to these AI systems.
68. This section provides guidance on transparency and explainability considerations arising from AI systems. Transparency is generally understood as providing accessible information about how an AI system is used, while explainability involves clarifying how the AI system arrives at its outcomes.

5.2 Explaining AI system outcomes

69. To prevent the market conduct and prudential risks described in the previous paragraphs, and in alignment with ICP 8 (Risk management) and ICP 19 (Conduct of business), supervisors should require that insurers are able to meaningfully explain the outcomes of the AI systems that they use. Such explanations are particularly important for those AI use cases that may have a material impact on consumers, solvency or satisfying legal requirements. Additionally, ICP 19.10 requires insurers "to handle claims in a timely, fair and transparent manner", so the transparency and explainability of claims decisions and claims dispute resolution influenced by AI systems are especially important here.
70. Meaningful explanations should be understood in the sense that they provide understandable, transparent and relevant insights into how the AI system makes decisions or predictions. There are several strategies and tools insurers can adopt to ensure their AI systems are explainable. For example, insurers could restrict deployment of AI systems to those that are simple and explainable, or restrict the use of complex AI systems to challenging and fine-tuning more traditional mathematical models. Alternatively, the deployment of complex AI systems could be conditional on the accompanying deployment of explainability tools such as Shapley values or Local Interpretable Model-Agnostic Explanations (LIME),[14] which can be employed to illustrate the influence of different variables on the outcomes of some AI systems, enhancing transparency and trust. However, even these state-of-the-art tools still have relevant limitations that need to

[13] See also the IAIS Application Paper on fair treatment of a wide range of consumers, expected to be published in July 2025.

[14] LIME and SHAP are two explainability techniques that aim to provide local explanations, ie an explanation about the behaviour of specific data points or regions in the input data (such as how they influence the output of the AI system).

be duly considered and documented by insurers, including their limitations in the context of GenAI.

71. For example, in insurance underwriting, for certain AI systems these tools can be used to explain why certain customers are offered different premiums. Insurers integrating SHapley Additive exPlanations (SHAP) values into their claims processing workflows can explain why certain claims were approved or denied. Furthermore, LIME can be used in underwriting to better explain risk assessments for insurance policies. By providing clear explanations of the factors/variables that influence risk scores, insurers can justify premium calculations to customers and regulators.
72. For highly complex AI systems (such as those incorporating a combination of unstructured data sets like images, video, audio and text), achieving an otherwise desirable level of explainability may not be possible. Where this is the case, insurers should consider adopting and documenting complementary governance measures such as the use of guardrails or human oversight. Additionally, where the risks from the AI system are high and/or the tools used to explain the model themselves have limitations, insurers could instead consider alternative simpler models.
73. In any case, insurers should work to improve the reliability of AI systems, including creating mechanisms to minimise risks in unforeseen situations. Where an AI system cannot provide sufficient confidence under new or unexpected conditions, insurers should design it to fail safely or escalate to human intervention. Systems should be designed to detect and address situations that fall outside their intended scope or where reliability is uncertain.

Table 4: Explaining AI system outcomes[15]

AI use case	Appropriate explainability approach	Possible explanation to customer
Underwriting	SHAP values to identify key factors	"Your premium was influenced primarily by your driving history (40% impact), vehicle type (30% impact) and location (20% impact)."
Claims processing	Decision trees for transparent logic	"Your claim was initially flagged for review due to the timing and nature of the incident, which matched patterns requiring additional verification."
Customer service chatbot	Confidence scoring with human escalation	"I'm 85% confident in my answer about your policy coverage. Would you like me to connect you with a human agent to verify?"

[15] While these explainability tools can provide relevant insights, as mentioned above, these tools still have relevant technical limitations that need to be duly considered and documented by insurers, and where relevant complemented with alternative governance and risk management measures.

5.3 Explanations adapted to the recipient stakeholders

74. Different stakeholders require different types of explanations, since not all stakeholders have the same technical knowledge or the same reason for seeking the explanation, nor do they require the same level of detail. A risk-based and proportionate approach should be taken; information needs will vary according to the risks associated with the use case and the relevance of the interaction. For instance, where an AI system is used for underwriting retail customers, it is appropriate that this is disclosed. However, where a model is used to process invoices for the finance team, the need for any disclosure is expected to be remote.
75. Consumers should be made aware if they are interacting with an AI system and be allowed to obtain assistance from a human if needed. Given that they may have limited knowledge of AI, consumers would require plain, simple and easy-to-understand information (for example, the use of visual aids and layman's terms) not involving the use of excessive technical language. An example is potentially providing, upon request, policyholders with a breakdown of the main factors that have influenced their premium calculations, such as age, driving history and geographic location to support explainability.
76. In contrast, other stakeholders such as auditors or supervisors will require more comprehensive and technical information about the AI system to allow them to perform an adequate supervisory review process (ICP 9). Such information could include, for example, information about how the data was collected, processes and post-processing methodologies, feature importance or the reasoning behind technical choices, including the governance and risk management measures put in place. Information should be sufficient to provide internal and external audit functions with the information they need to make a proper assessment of the extent to which policies have been effectively followed.
77. Furthermore, it is also important to respect trade secrets and confidentiality obligations, and therefore the information to be provided may vary from one use case to another. For example, with certain use cases such as fraud detection, insurers may not be able to disclose detailed information to consumers about their practices. By also taking into account intellectual property considerations, as noted in Section 3.5 above, insurers should obtain adequate information and reassurances from third-party service providers regarding the AI systems that they purchase from them.

6 Fairness, ethics and redress

6.1 Introduction

78. AI systems allow a wide range of data sets to be consolidated and analysed to support decision making. Insurers therefore need effective data governance processes. AI systems can be susceptible to biases and other stereotypes present in training and secondary data sources. Bias could inadvertently be programmed into AI system protocols, leading to unfair or discriminatory decisions if not properly managed. Furthermore, AI systems can be used to manipulate or exploit consumers' behavioural biases, such as their willingness to pay or propensity to shop around at the renewal stage of the contract. While these pricing practices may exist with and without the use of AI systems, they can potentially lead to unfair or unethical outcomes if there are no adequate governance and risk management measures in place.
79. Protection against unlawful discrimination is enshrined in international treaties and jurisdictional legal systems. It is therefore important that the use of AI systems does not diminish such

protections. However, it is important to differentiate between unlawful discrimination versus lawful risk differentiation and risk-based pricing, where the decision of whether to provide coverage and what premium to charge a customer is connected to the customer's level of risk.[16]

80. ICP 19 requires that insurers and intermediaries treat customers fairly both before a contract is entered into and through to the point at which all obligations under a contract have been satisfied. Treating customers fairly applies where AI models are being used. This requirement promotes fair consumer outcomes at each stage of the product life cycle (further elaborated in ICP 19.0.2)[17] and encompasses concepts such as "ethical behaviour, acting in good faith and the prohibition of abusive practices" (ICP 19.0.3).
81. Insurers should adopt a fairness-by-design approach, embedding fairness into AI governance and risk management. Building on Section 5, which includes relevant measures from a fairness perspective, this section elaborates on further key fairness and ethical considerations arising from AI systems and proposes ways to address them. As with other dimensions of conduct of business, what is considered to be fair or ethical is closely linked with "jurisdictions' tradition, culture, legal regime and the degree of development of the insurance sector" (ICP 19.0.3).

6.2 Data management in the context of fairness

82. AI systems and their outcomes rely extensively on data; thus, biases or inaccuracies in the data sets used to train the AI system may, without appropriate controls, be reproduced in the outcome. It is important therefore that data sets used for training AI models be accurate, complete and representative of the customer segment being served and that data use is monitored to mitigate bias. Section 3.6 and the Annex highlight the importance of record keeping, the role of data in assessing the model's robustness, and data security, respectively. The points below provide additional considerations for supervisors to assess whether insurers have adequate data management processes throughout the AI system life cycle in order to promote fairness in how the data is used and to mitigate errors and biases that could emerge during data collection, processing and application:
    - *Data collection:* Carefully select diverse and relevant data sources that are appropriate for the intended use of the AI systems.
    - *Data preparation:* After collection, data should be processed to ensure accuracy (no material errors and free of bias) and completeness (representative of the population and sufficient historical information). This involves exploring and cleaning the data to remove duplicates

[16] See the IAIS Application Paper on how to achieve fair treatment for a wide range of consumers, which further explores this distinction (Sections 2.1 and 2.2).
[17] ICP 19.0.2 notes that fair treatment of customers encompasses achieving outcomes such as:

- Developing, marketing and selling products in a way that pays due regard to the interests and needs of customers;
- Providing customers with information before, during and after the point of sale that is accurate, clear and not misleading;
- Minimising the risk of sales which are not appropriate to customers' interests and needs;
- Ensuring that any advice given is of a high quality;
- Dealing with customer claims, complaints and disputes in a fair and timely manner; and
- Protecting the privacy of information obtained from customers.

and invalid data and complete missing values. Traceability in data transformation is crucial to monitor its impact on AI systems.

- *Post-processing:* The outcomes of the AI system should be assessed for data quality and potential discriminatory biases (see further below). A correction/verification loop is essential for maintaining data integrity.

83. The insurer's data management processes should govern against using customer data in an unfair manner (ICP 19.12.7). For example, when data is processed by an AI system for non-risk-based pricing practices, where allowed at all, they should have in place governance and risk management processes that prevent the unfair treatment of consumers, for example by defining adequate thresholds or guardrails of consumers with similar risk profiles from different distribution channels. Moreover, there should also be policies and processes for "ensuring that customers have a right to access and, if needed, to correct data collected and used by insurers and intermediaries" (ICP 19.12.7).

6.3 Inferred causal relations in an AI system

84. Model training, whether in AI or non-AI defined systems, involves using historically identified correlations to infer causality[18]; however, inferences are not facts, as history is only one input to insurability. The additional complexity presented by AI systems relates to the significant increase in volume and variety of data analysed (often a combination of primary and secondary data sources) and the complexity of the underlying algorithms, such that the correlations (often non-linear and multivariable), and implied inferences of causality they make, can be difficult to identify. In this context, it is useful to reiterate that identified correlations do not necessarily imply causation.[19]

85. As part of the appropriate policies and processes to prevent unfair use of data (ICP 19.12.7), for certain high-risk AI use cases insurers should establish a process to regularly extract and document the implied AI system inferences (and hence implied causal relationships) in a clear and transparent manner such that rational explanations can be provided. Such documentation should enable effective challenge and discussion on whether the implied causal relationships are in line with expectations and the insurer's strategic objectives (for example, the extent to which predictions from an AI system infer causality based on identified correlations that reflect historic societal biases). Such documentation should support Senior Management and underwriters in assessing the extent to which decisions are risk based and consistent with legal and regulatory obligations.

86. There should also be policies and processes in place so customer data is not abused to cause unlawful discrimination (ICP 19.12.7). In this respect, insurers should carefully consider the use of proxy variables, especially in pricing and underwriting practices.

[18] Causal models are a specialised subset of modelling approaches designed to infer causation from data.

[19] For example, ice cream sales and shark attacks in the United States are highly correlated; however, this does not mean that eating ice cream causes shark attacks. The more likely explanation is that people consume more ice cream and swim in the ocean when it is warmer outside, leading to the correlation (correlation ≠ causation).

6.4 Monitoring the outcomes of AI systems

87. Traditionally, there has been a greater emphasis on ex ante governance processes, such as those described in Section 6.2 above, to minimise the likelihood of discriminatory outcomes. However, the deployment of AI systems may require a greater emphasis on processes designed to monitor outcomes (ICP 19.0.2).
88. For example, some AI systems such as neural networks or DL algorithms are capable of capturing non-linear multivariable dependencies amongst the training data that may replicate protected characteristics (eg multivariable dependencies between address, job and shopping habits may closely correlate with a customer's ethnicity or gender). These dependencies may go undetected by the human programmer of the AI system due to the limited explainability of the AI systems.
89. Furthermore, AI systems are often trained with data sets provided by third parties (sometimes referred to as secondary data). With such data sets, due to intellectual property considerations, it is often difficult or challenging for the insurer to identify and thoroughly assess the data processing methodologies that have been used by the provider. Examples of such cases include credit scores provided by credit rating agencies or, more recently, foundation models (including LLMs) underlying GenAI systems.
90. ICP 19.12.7 requires that "the supervisor should not allow insurers and intermediaries to use customer information that they collect and hold in a manner that results in unfair treatment". Therefore, the policies and processes of the insurer should embed appropriate governance and risk management measures according to the AI use case, such as using more explainable AI systems and using fairness metrics to assess model outcomes in high-risk AI use cases. Provided it is legally permitted in the respective jurisdiction, this last approach may involve the collection of protected information from customers or the use of aggregated population data at the postcode level obtained from the census, municipalities, tax authorities or other relevant agencies. The insurer's policies and processes should provide for documentation of the outputs of AI systems, as well as for documentation of the results of any fairness testing performed on those outputs. Supervisors should consider whether to require insurers to keep an inventory of models with varied levels of information depending on the complexity of the AI system and its use case. Such an approach should be proportionate to the risks of the AI system. Supervisors will be able to check the accuracy and completeness of the model inventory.
91. Some examples of fairness (functional correctness) metrics are provided in the Annex. The use of different fairness metrics may vary for different AI use cases. They help with monitoring model outcomes and, subsequently, introducing changes in the model to obtain the desired fairness output. Insurers may consider developing fairness dashboards that monitor key metrics across different customer segments over time, enabling early detection of emerging disparities that may require intervention.

6.5 Adequate redress mechanisms for claims and complaints

92. When AI systems (regardless of their level of complexity and explainability) are used in decision-making processes, disputes can arise between the affected stakeholders. For example, a consumer may want to understand why their application for an insurance product has been rejected or why their compensation for a claim is not as much as they were expecting. In line with ICPs 19.10 and 19.11, supervisors should require that insurers have in place effective, fair and transparent redress mechanisms, both for claims and complaints disputes. In this context, for high-risk AI use cases, it is particularly important for insurers to give meaningful explanations

on determinative factors in claims or complaints resolution (for example, by using more explainable AI systems). This will enable those adversely affected by an AI system to challenge its output. As previously noted, insurers are expected to remain accountable for appropriately explaining decision-making that affects consumers regardless of the tools or models used.

93. Part of this redress mechanism should include the ability for a consumer to update, supplement or correct information and data from sources that are used in the AI systems. Additionally, it should include the ability to seek human intervention. This will allow consumers to challenge and update information from third-party data sources as well as information generated by the insurer. This is consistent with best practice policies on data protection. In order to make these changes, it is possible that human intervention will be required.

6.6 Societal impacts of granular risk pricing

94. The use of AI-driven, risk-based pricing in insurance has the potential to enhance financial inclusion by enabling the underwriting of risks that were previously considered uninsurable due to limited data availability. For example, improvements in health data and AI capabilities may allow insurers to offer coverage with specific conditions rather than declining coverage altogether. While the application of large data sets and advanced AI techniques can support more granular and actuarially sound pricing, it may also lead to increased premium differentiation across population segments. Nevertheless, such developments could contribute to a more efficient and responsive insurance system overall.

95. From a societal perspective, there is a risk that highly granular AI-enabled pricing could result in unaffordable premiums for certain vulnerable groups, such as low-income households or minority communities. To address this concern, insurers and supervisors may consider implementing mitigating measures. These could include monitoring and addressing premium disparities, assessing the broader socio-economic impacts of pricing practices, and restricting the use of risk factors deemed unfair or discriminatory. Addressing potential societal impacts of granular risk pricing requires proactive engagement among stakeholders, including AI developers, insurers, and consumer representatives. Advisory panels or working groups can focus on AI ethics and fairness while adhering to competition laws. Supervisors play a vital role in addressing these issues by analysing market trends, engaging with insurers, and balancing legitimate underwriting practices with financial inclusion to reduce protection gaps.

Annex: Examples from IAIS Members

Supervisors are already taking steps to address risks from AI. This section sets out some illustrative examples of actions supervisors are taking. Examples here are jurisdiction specific and therefore should be seen within the specific legal and regulatory contexts in which they operate.

Proportionality and risk-based supervision

Classifying AI systems and applying proportionality: an example from the EU

The AI Act[20] applies to all sectors of the European economy and aims to ensure a high level of protection for the fundamental rights, health and safety of AI systems. The AI Act follows a risk-based approach and creates a framework for classifying AI systems according to different risk levels:

- Unacceptable risks: AI risks that are deemed to be unacceptable are prohibited and should not be brought into the market.
- High risks: Providers and users of high-risk AI systems will need to comply with comprehensive governance and risk management requirements. In the insurance sector, the AI Act identifies as high-risk those AI systems intended to be used for risk assessment and pricing in relation to natural persons in the case of life and health insurance.
- Limited and minimal risks: This category encompasses the majority of AI use cases and sets out minimum transparency requirements, a general AI literacy requirement and the development of voluntary codes of conduct.

Due to their specific nature and complexity, including their ability to perform a wide variety of different tasks and use cases, specific rules are also established for the so-called general purpose AI systems (eg LLMs).

Insurance sector legislation continues to apply to all AI use cases in insurance, regardless of their qualification under the AI Act. To address potential overlaps, the AI Act introduces limited derogations applicable to undertakings subject to Solvency II.

Application of proportionality: an example from the NAIC

In the United States, acting through the National Association of Insurance Commissioners (NAIC), state insurance regulators adopted a Model Bulletin on the Use of Artificial Intelligence Systems by Insurers[21] on 4 December 2023.

The bulletin indicates that the controls and processes that an insurer adopts should be reflective of, and commensurate with, the degree and nature of risk posed to consumers. To this extent, the

[20] See EU, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 13 June 2024.

[21] See NAIC, NAIC Model Bulletin: Use of Artificial Intelligence Systems by Insurers, 13 Oct 2023.

bulletin provides guidelines that align with the NAIC principles on AI[22] for assessing the risks of AI systems.

The adoption of proportionate governance, risk management controls and internal audit functions aligning to the level of risk should be developed to avoid violating the Unfair Trade Practices Acts and other applicable laws and regulations.

Data governance and record keeping

In 2021, EIOPA created a Consultative Expert Group on Digital Ethics, which developed a report on AI governance principles[23] aimed at guiding European insurers in the development and use of ethical and trustworthy AI systems. The group proposed a set of record-keeping practices for high-risk AI systems. The table below includes these, plus other examples.

Record	Description
Reasons for using AI	Explanation of the business objective/task pursued by using AI and its consistency with corporate strategies/objectives. Explanation of how these objectives were implemented into the AI system (ie what are the goals prescribed in the AI system). This would help reduce misuse of the AI system and enable its audit and independent review.
Integration into IT infrastructure	Description of how the system is integrated in the current IT system of the organisation and document any significant changes that could eventually take place.
Staff involved in the design and implementation of the AI system	Identify all the roles and responsibilities of the staff involved in the design and implementation of the AI system as well as their training needs. This supports achieving accountability of the responsible persons.
Data collection	Document how the ground truth[24] was built including how consideration was given to identifying and removing potential bias in the data. This would include

[22] Materials - Innovation and Technology (EX) Task Force: https://content.naic.org/sites/default/files/inline-files/AI%20principles%20as%20Adopted%20by%20the%20TF_0807.pdf.

[23] See EIOPA, Artificial intelligence governance principles: towards ethical and trustworthy artificial intelligence in the European insurance sector, 2021.

[24] Ground truth is information that is known to be real or true, provided by direct observation and measurement as opposed to information provided by inference (2021).

	explaining how input data was selected, collected and labelled.
Data preparation	Records of the data used for training the AI system (ie the variables with their respective domain range). This would include defining the construction of training, test and prediction data set. For built (engineered) features, records should exist on how the feature was built and the associated intention.
Data post processing	Description of processes in place to operationalise the use of data and to achieve continuous improvement (including addressing potential bias). Records should specify the timing and frequency of data improvement actions.
Technical choices/arbitration	Document why a specific type of AI algorithm was chosen and not others, as well as the associated libraries with exact references. The limitation/constraints of the AI system should be documented and how they are being optimised alongside their supporting rationale. Ethical, transparency and explainability trade-offs that may apply together with their rationale should also be recorded.
Code and data	Record the code used to build any AI system that is in a "live" environment. Additionally, for high risk applications, insurance firms should record the training data used to build the AI system and all the associated hyper parameters, including pseudo-random seeds.
Model performance	Explanations should include, inter alia, how performance is measured (key performance indicators) and what level of performance is deemed satisfactory, including scenario analysis and timing and frequency of reviews and/or retraining of the model. Ethical, transparency and explainability trade-offs that may apply together with their rationale should also be recorded.
Model security	Describe (or make reference to) mechanisms in place to ensure the model is protected from outside attacks and more subtle attempts to manipulate data or algorithms themselves: how robust is the model to

	manipulation attacks (especially important in auto ML models)?
Ethics and trustworthy assessment	Description of the AI use case impact assessment (ie the potential impact on consumers and/or insurance firms of the concrete AI use case). Explain how the governance measures put in place throughout the AI systems life cycle address the risks included in the AI use case impact assessment and ensure ethical and trustworthy AI systems. Records should include individuals and groups that are considered to be at risk of being systematically disadvantaged by the system, including the potential harms and benefits, and the fairness objectives of the system and associated fairness metrics. The records should show in practice how these groups are impacted.

AI safety and security

Code of practice for AI cyber security: an example from the UK[25]

In May 2024, the UK government issued a public consultation on the cyber security of AI, which includes a voluntary Code of Practice that emphasises a secure-by-design approach throughout the life cycle of AI technologies. The Code of Practice principles are:

Secure design

- Raise staff awareness of threats and risks;
- Design your system for security as well as functionality and performance;
- Model the threats to your system; and
- Ensure decisions on user interactions are informed by AI-specific risks.

Secure development

- Identify, track and protect your assets;
- Secure your infrastructure;
- Secure your supply chain;
- Document your data, models and prompts; and
- Conduct appropriate testing and evaluation.

[25] See UK Department for Science, Innovation and Technology, Cyber security codes of practice, 15 May 2024.

Secure deployment

- Communication and processes associated with end users.

Secure maintenance

- Maintain regular security updates for AI model and systems.
- Monitor your system's behaviour.

Cyber risks associated with GenAI: an example from the MAS[26]

In July 2024, MAS issued an information paper on the cyber risks associated with GenAI, which include threats and risks on GenAI deployments (namely, unauthorised information disclosure and data leakage), as well as GenAI model and output manipulation.

Data leakage

Risks

- Upload of sensitive data by staff into public GenAI tools; and
- Prompt injection attacks or jailbreak attacks.

Possible mitigation measures

- Establish user policies and conduct employee awareness campaigns on security best practices in relation to GenAI usage;
- Adopt security best practices when developing in-house GenAI models, such as implementing security-by-design approach and secure coding, performing vulnerability assessments and security testing;
- Perform proper due diligence when using third-party or open-source GenAI solutions; and
- Implement data loss prevention and firewalls for GenAI models.

Model/output manipulation

Risk

- Threat actors can introduce malicious or inaccurate data, for example through data poisoning attacks, to manipulate the GenAI models and their outputs. This can take place during the training stage or while using the models.

Possible mitigation measures

- Put in place proper GenAI model and data governance;
- Ensure robust access controls to the GenAI training data and foundation model;
- Implement continuous monitoring and validation of GenAI models;
- Incorporate contingency measures for GenAI solutions into business continuity plans; and
- Participate in information sharing to identify issues related to GenAI model deployment.

[26] See MAS, Cyber Risks Associated with Generative Artificial Intelligence, July 2024.

Considerations for AI system transparency

Disclosure of credit scores: an example from the United States

When considering disclosures that could be made to consumers about how decisions are made using AI systems, existing frameworks, such as that for credit scoring, could provide some useful parallels. For instance, the US Fair Credit Reporting Act sets out requirements on statements consumers have a right to receive based on how their data feeds into credit scoring. The statement includes:

- The current credit score of the consumer or the most recent credit score of the consumer that was previously calculated by the credit reporting agency for a purpose related to the extension of credit;
- The range of possible credit scores under the model used;
- All of the key factors that adversely affected the credit score of the consumer in the model used, the total number of which shall not exceed four;
- The date on which the credit score was created; and
- The name of the person or entity that provided the credit score or credit file upon which the credit score was created.

The term "key factors" means all relevant elements or reasons adversely affecting the credit score for the particular individual, listed in order of their importance based on their effect on the credit score. Supervisors may want to consider what elements here may be applicable to disclosures about the use of AI systems.

Ensuring communications to users are appropriate: an example from UK actuarial standards

In October 2024, the Financial Reporting Council published[27] updated guidance to support practitioners in complying with technical actuarial standards when using models that include AI and ML techniques. The council considered risks that may be increased by the use of AI and ML techniques and how to address those risks. Four examples were included in the guidance, covering model bias, understanding and communication, governance and stability.

The understanding and communication example includes a number of actions taken to understand and explain the models employing AI and ML techniques used for a piece of actuarial work:

- Using a range of techniques that help show the relationships between input variables and output variables. This includes understanding and considering limitations of these techniques.
- Considering both the intrinsic understandability of the proposed models and the explainability of the models based on the use of techniques.

[27] See FRC publishes updated actuarial guidance on the use of AI and Machine Learning.

- Balancing explainability with other factors, including accuracy when choosing between models, taking into account the intended user and their level of technical knowledge.
- In the communications to the intended user, providing an outline of how the model works and an explanation of key judgments made. This includes reporting on how the model responds to changes in key input variables, as shown through the application of techniques to increase explainability, and any limitations of these techniques that are considered material to the decision being made.

Information from third-party service providers

Expectations around third-party service providers: an example from the NAIC

The NAIC Model Bulletin provides guidance on the governance and risk management measures to be adopted by insurers using AI systems. Specifically concerning outsourcing from third parties, the AI system bulletin sets forth the following expectations:

"Each AIS system governance program should address the Insurer's process for acquiring, using, or relying on (i) third party data to develop AI Systems; and (ii) AI Systems developed by a third party, which may include, as appropriate, the establishment of standards, policies, procedures, and protocols relating to the following considerations:

4.1 Due diligence and the methods employed by the Insurer to assess the third party and its data or AI Systems acquired from the third party to ensure that decisions made or supported from such AI Systems that could lead to Adverse Consumer Outcomes will meet the legal standards imposed on the Insurer itself.

4.2 Where appropriate and available, the inclusion of terms in contracts with third parties that:

a) Provide audit rights and/or entitle the Insurer to receive audit reports by qualified auditing entities.

b) Require the third party to cooperate with the Insurer with regard to regulatory inquiries and investigations related to the Insurer's use of the third-party's product or services.

4.3 The performance of contractual rights regarding audits and/or other activities to confirm the third-party's compliance with contractual and, where applicable, regulatory requirements."

Monitoring outcomes from AI systems

Ensuring compliance with local regulations: an example from the NYS DFS

In July 2024, the New York State Department of Financial Services (NYS DFS) published a Circular Letter on the Use of AI Systems and External Consumer Data and Information Sources

in Insurance Underwriting and Pricing.[28] The Circular Letter outlines the key governance and risk management measures that insurers are expected to implement to ensure compliance with local regulations.

Amongst other measures, insurers are encouraged to use multiple statistical metrics in evaluating data and model outputs to ensure a comprehensive understanding and assessment, including the following:

- *Adverse impact ratio*: Analysing the rates of favourable outcomes between protected classes and control groups to identify any disparities.
- *Denials odds ratios*: Computing the odds of adverse decisions for protected classes compared with control groups.
- *Marginal effects*: Assessing the effect of a marginal change in a predictive variable on the likelihood of unfavourable outcomes, particularly for members of protected classes.
- *Standardised mean differences*: Measuring the difference in average outcomes between protected classes and control groups.
- *Z-tests and T-tests*: Conducting statistical tests to ascertain whether differences in outcomes between protected classes and control groups are statistically significant.
- *Drivers of disparity*: Identifying variables in AI systems that cause differences in outcomes for protected classes relative to control groups. These drivers can be quantitatively computed or estimated using various methods, such as sensitivity analysis, Shapley values, regression coefficients or other suitable explanatory techniques.

HKIA's supervision on chatbots and AI

Recognising the potential impact of AI-powered chatbots on the insurance sector, the Hong Kong Insurance Authority (HKIA) published user guides in its May 2023 *Conduct in Focus* series.[29] These guides outline key considerations and perspectives on their implementation under the "regulated activities" regime. The user guides include considerations such as:

- Legal challenges such as copyright issues surrounding chatbot-generated content and the fact that accountability for their outputs should rest on the insurer or intermediary deploying the chatbots.
- Cyber security, confidentiality and personal data implications, especially as these technologies can be misused for malicious purposes.
- The importance of risk evaluation, comprehensive testing before deployment, and adherence to guidelines on ERM, outsourcing and cyber security.

[28] See NYS DFS, Insurance Circular Letter No 7 RE: Use of Artificial Intelligence Systems and External Consumer Data and Information Sources in Insurance Underwriting and Pricing, July 2024.

[29] See HKIA, "Chatting about Chatbots and AI", Conduct in Focus, May 2023.

- Clear disclosure would need to be made as to the chatbot's limitations, how it should be used, the data set it is trained on and how that data is stored and used and how long it is kept. Adequate risk mitigation, ongoing monitoring, reporting controls and contingency plans would also need to be in place throughout its deployment.

- It is crucial for insurers and insurance intermediaries using AI to uphold principles of fair customer treatment, honesty and integrity, acting in the customer's best interests and enabling fully informed customer decisions.

The HKIA is also exploring the development of a comprehensive regulatory framework that promotes the fair, transparent and ethical use of AI in the insurance industry while adequately addressing concerns such as algorithmic bias and personal data leakage.